2018 10th International Conference on Cyber Conflict
CyCon X: Maximising Effects
T. Minárik, R. Jakschis, L. Lindström (Eds.)
2018 © NATO CCD COE Publications, Tallinn

Rethinking the Data Wheel: Automating Open-Access, Public Data on Cyber Conflict

Christopher Whyte
Assistant Professor
Virginia Commonwealth University

Brandon Valeriano
Donald Bren Chair of Armed Politics
Marine Corps University

Benjamin Jensen
Associate Professor
Marine Corps University

Ryan Maness
Assistant Professor
Naval Postgraduate School

Abstract: To date, researchers studying cyber conflict through publicly available information sources have either selected on the actor or selected on the intrusion method when coding events. Both approaches lead to distinct challenges when it comes to result validation and the avoidance of selection bias. This article describes prospects for open-source, public data collection for cyber security events. We present an initial data collection and analysis effort of interstate cyber conflict incidents involving the United States as a pilot study. Using a tailored collection of more than 155,000 documents from print-only media sources, we describe a method to process data, parse document elements, and populate an event dataset. Human coders are then tasked with validation of incident information, after which the search code is updated to ensure greater accuracy in subsequent runs. In the study, the data produced are compared with previously available data on cyber conflict involving the United States. We demonstrate that the method can effectively capture and describe cyber conflict incidents for researchers to study in a broad range of research efforts. Moreover, this method captures greater granularity within cyber conflict episodes, which are inherently multi-faceted. This approach to cyber conflict analysis carries with it several distinct advantages over alternative research designs, in that it promises to produce significantly larger amounts of pertinent metadata than might otherwise be possible.

1. INTRODUCTION

Researchers analyzing the scope and scale of global cyber conflict face significant data collection challenges. In particular, the process of determining who is responsible for observed cyber incidents that are often covert by design produces research constraints for researchers seeking to describe modern competition, conflict and confrontation empirically (Gartzke and Lindsay, 2015; Rid and Buchanan, 2015). How can researchers systematically study cyber incidents globally and document recurrent patterns and trends, given inherent restrictions on coding what are essentially covert operations?

Such challenges are pressing for scholars and practitioners alike insofar as both aim to develop a sophisticated body of knowledge regarding the drivers, determinants, and effects of conflict waged via networked information and communications technologies (ICT). To date, the cyber security field tends to rely on thin case study descriptions of cyber incidents, using crucial cases to make inferences about actor motivation and the larger context of the cyber conflict, as well as using deductive reasoning to produce a foundation of theoretical knowledge regarding cyber conflict. For example, major work on the Stuxnet attack tends to take this form, with scholars debating the efficacy and larger implications of the series of espionage and degradation intrusions launched by multiple states against Iranian targets (Lindsay, 2013; Slayton, 2017; Kello, 2017). With respect to deductive reasoning, major studies use a series of anecdotal examples to work through a series of logical claims about cyber deterrence and crisis escalation in cyberspace, even including paralyzing cyber first-strikes and offensive action (Libicki, 2012; Gompert and Libicki, 2014; Whyte, 2016; Nye, 2017). Despite its classified nature, most intelligence analysis of cyber events likely replicates these methods. Faced with a poverty of data, analysts and scholars alike use individual incidents and deductive reasoning to illuminate emerging threats and opportunities in cyberspace.

To date, research that systematically collects data on cyber incidents is scarce. Outside of work on cyber rivalry and limited studies of denial of service attacks within a conflict setting, both of which limit the sample under investigation, most of the cyber security literature lacks large databases and robust samples (Valeriano and Maness, 2014; Valeriano and Maness, 2015; Kostyuk and Zhukov, 2017; Whyte, 2017; Valeriano, Jensen, and Maness, 2018). The absence of large datasets limits the development of inductive meta-theories about cyber conflict. Policy makers and scholars cannot determine whether an intrusion event is an isolated and insignificant incident, or consistent with a larger correlate of cyber conflict, without understanding the true scope of cyber interactions.

For scholars interested in the cyber domain, assessment of information derived from publicly-available outlets is an option that is as attractive as it is problematic. The capture and treatment of massive amounts of published data pertaining to cyber conflict promises a unique resource for those seeking to assess the context of cyber security engagements. Nevertheless, such approaches often garner broad criticisms pertaining to generalizability and methodology. If much of what constitutes cyber conflict is covert, how can data produced from information found in the public sphere offer researchers the opportunity to generalize? Even if that hurdle were to be cleared, how can researchers reconcile attribution challenges in determining the sources, targets and technical shape of varied cyber interactions? Without some notion of reliability as a measurement of the value of such information, open source data efforts are likely to run into serious problems.

This article addresses the data challenge at the core of cyber security. First, we address the utility of open-source data collection on cyber conflict processes for scholars and practitioners alike. In addition to being the most promising route available for academic researchers to develop a robust knowledge foundation from which to undertake sophisticated analyses, assessing open access materials both allows researchers to look at the context of cyber conflict and provides opportunities for use of advanced analytic methods that can parse signal from immense noise. Second, we describe an approach – commonly found in research on political violence, and in recent efforts to build comprehensive conflict event data – for producing cyber conflict data that draws from public-facing information sources and allows the researcher to address validation shortcomings inherent to such an approach. Then, we demonstrate the value of this approach by employing a tailored collection of more than 155,000 documents from print media sources in the United States, in order to produce data on interstate cyber interactions across a two-year period. This approach performs on par with data previously produced via traditional collection approaches and, insofar as different elements of episodes are captured, produces a more granular picture than has been produced in prior large-N work on cyber conflict. Likewise, opportunities to enrich such data via additional treatment of surrounding text and linked documentation promise further value to researchers seeking to understand the sociopolitical context of cyber conflict (Schrodt, Beieler, and Idris, 2014).

The article proceeds in five sections. First, it considers the state of cyber conflict data production, describes the few attempts that have been made to date to produce systematic accounts of warfare conducted online, and outlines enduring challenges. Then, we make a case for the clear utility of data produced from publicly-available information sources. Third, we describe the requirements for robust, replicable efforts to develop such data resources for scholarly use, before demonstrating this via the presentation of two years' worth of event data on interstate cyber conflict involving

the United States. We conclude with a discussion of the implications of our arguments, and a demonstration for both researchers and policymakers as well as practitioners.

2. CYBER CONFLICT DATA: PRIOR EFFORTS AND ENDURING CHALLENGES

The incidence of cyber conflict dates back to the early 1980s with episodes such as the Farewell incident, in which the CIA targeted KGB technology transfer programs, and the Cuckoo's Egg hack-back, in which private network operators identified Soviet operatives (Stoll, 1988; Healey and Grindal, 2013). In spite of this, systematic and comprehensive resources describing cyber conflict incidents are virtually non-existent. Major political science efforts to catalogue different forms of interstate conflict and political violence fail to include cyber actions, either owing to their ambiguous origins or to difficulty attributing the incident. Stuxnet, for example, although a crucial case in descriptive treatments, is often not represented in major databases due to attribution issues, difficulty dating the start and end of the incident, and the question of whether it was the United States or Israel that launched the action (Radford, 2016).

This general lack of focus on cyber conflict issues in the context of broader efforts to record and problematize international security dynamics is troubling for a number of reasons. Foremost among these is the fact that there is arguably a consensus among political scientists that cyber instruments work as adjunct modifiers – essentially force multipliers – of conventional and asymmetric warfare (Gartzke, 2013, Valeriano et al., 2018). This suggests that cataloguing cyber incidents is useful not only as a means of assessing conflict restricted to that domain, but also as a means of understanding a critical variable in broader conflict processes. Without better understanding of the nature of cyber conflict, scholars and security practitioners of all stripes are (and will be) hard pressed to describe accurately how digital actions and possibilities intersect with existing mechanisms of human interaction. Indeed, without such a development, it is likely that we inject bias – from using data obtained only from select stakeholders or employing methods that misunderstand the significance of different actors – into our continued efforts to construct knowledge of macro global security processes.

The main reason that no comprehensive data resource to describe cyber conflict exists is that the attribution of cyber incidents is not always feasible (Rid and Buchanan, 2015; Lindsay, 2015). This is true on two fronts. Firstly, the method is covert: while there are often observable outputs of cyber conflicts, where victims (or, in rare instances, observers) report on incidents or attackers broadcast their involvement, this is not always true. Indeed, anecdotal evidence and simple recognition of the scope of the domain to be canvassed by researchers suggests that this is true only infrequently.

Bound up in this problem is the manner in which the digital world operates. Whereas with other forms of conflict – terrorist attacks, for instance – it may be possible to adjudicate reasonably on the frequency of otherwise invisible attacks based on knowledge of past actions, analytic breakdown of capabilities, or journalistic efforts to validate rumor, the same is not generally accepted in cyberspace. Even where indicators suggest the existence of incidents to researchers, validation usually requires the cooperation of victims or infrastructural stakeholders (i.e. backbone operators or non-backbone ISPs). Thus, particularly where relevant actors are motivated by the possibility of reputational, financial or political costs, confirmation of the full scope of cyber conflict is difficult for those operating in the public domain.

Added to these challenges are the dual problems of bounding scale and controlling for negative cases. With respect to scale, a successful cyber operation might involve thousands of individual intrusion incidents. For example, spear phishing campaigns that resulted in the compromising of the German Bundestag and, more recently, the U.S. Senate, involved hundreds of e-mails sent to unsuspecting elected officials and staffers.[1] Does each e-mail constitute an individual cyber intrusion, or can researchers include them all as one campaign? Regarding negative cases, researchers must acknowledge the fact that cyber security firms, journalists, and governments tend to report only successful intrusions or attempts that nevertheless cause at least some measure of disruption (Brodsky, 2008). Unsuccessful intrusions, which likely are significantly larger by count, are thus under-reported.

Similarly, the second facet of the reporting problem lies with the value of information that can be obtained. Though such challenges are often surmountable, as we describe below, it is certainly true that gathering enough detail on a given incident to allow sociopolitical attribution is possible but difficult. Despite the clear imperative social scientists have to use any and all information available in attempting to understand the world around them, efforts to understand cyber phenomena better regularly run into criticism, as operating in a covert domain will generate no observable data (Lewis, 2002). This point fundamentally misunderstands the meaning of covert action, however, which implies a difficulty in determining responsibility, but not whether or not the event occurred.

Datasets are routinely released in the broad international relations field cataloguing all manner of security phenomena.[2] Among these, a small number are broadly focused on conflict with a relatively unlimited remit. Rather than focus solely on the efforts of terrorist non-state actors, insurgent movements, social activists or state militaries, such data collection efforts aim to catalogue the full spectrum of conflictual incidents

1 See *inter alia* http://www.zeit.de/digital/2017-05/cyberattack-bundestag-angela-merkel-fancy-bear-hacker-russia and http://thehill.com/policy/cybersecurity/368671-russia-linked-hackers-targeting-us-senate.

2 See, for instance, the Militarized Interstate Dispute dataset (http://cow.dss.ucdavis.edu/data-sets/MIDs) at the Correlates of War project, the International Crisis Behavior project (https://sites.duke.edu/icbdata/) and the Uppsala Conflict Data Program (http://ucdp.uu.se).

around the world. Over the past few years, such efforts have rapidly become more sophisticated. Efforts like Phoenix[3] and the Integrated Conflict Early Warning System (ICEWS)[4] provided extremely granular information on the nature of security events worldwide using a series of automated data scraping, parsing and treating methods, often in tandem with human validation inputs. Such approaches constitute the new normal for political science researchers in terms of the resources being made available to study international conflict. And yet, these macro efforts to describe global security matters do not systematically aim to capture all manner of cyber incidents (though they may include individual, prominent events) as part of their approach. This is possibly because the various attack chain elements that constitute the wide array of techniques of interest to cyber conflict scholars are not obviously conflictual in nature, and thus present a challenge when determining inclusion.

To date, there is only one dataset that accounts for all actors, states, and regions in the world available to scholars interested in the contours of global cyber conflict. The Dyadic Cyber Incident and Dispute dataset (DCID) describes interstate cyber conflict over more than fifteen years and employs a Correlates of War (CoW)-style coding scheme to describe the character of cyber warfare campaigns among rival states. The authors of DCID, Valeriano and Maness (2014, 2015), include a range of information on the type of instruments involved in observed cyber events, the impact of such events, and more. The data collected originates from publicly-available descriptions of cyber conflict incidents, including news stories, industry and government reporting, and expert testimony. Nevertheless, as the authors freely admit and others note (Radford, 2016), DCID was designed as an initial effort to scope the cyber conflict domain by selecting on rival states most likely to engage in cyber conflict. It is not aimed at the production of cross-domain conflict data, and does not draw from the universe of possible information on cyber incidents in a comprehensive sense. While outputs of the project might describe contours of cyber conflict between rival actors, any comprehensive effort to produce cyber conflict data must inevitably drop such selection parameters in order to ensure generalizability. Thus, the need to address the role of future open source data collection on cyber conflict is twofold, insofar as researchers must grapple with *both* absent resources and limited foundational efforts from which to begin their investigations.

Briefly, the data collection approach we describe below addresses these dual needs and goes a step further than previous social science projects. We rethink prior approaches to data collection in line with work undertaken in political violence and terrorism research programs, and expand beyond a limited focus rival states. In doing so, we provide for reliability checks that have been absent – or hard to effect – in past efforts, and argue that sophisticated data collection in this vein must turn to human reliability checkers for all machine learning processes. The result would be a dataset

3 See http://openeventdata.org/datasets/OEDA.datasets.php.
4 See https://dataverse.harvard.edu/dataverse/icews.

both large and relatively free of the errors common to other large event databases, such as ICEWS (Boschee et al., 2015) or IDEA (King and Lowe, 2003). Part of the reason we argue that this will be the case is the fact that projects like ICEWS and IDEA aim to capture all events between all actors annually. A cyber conflict effort would include a significantly reduced scope of inquiry, and would make the parsing of signal from noise a more feasible task. In short, though we cede the point that there are limitations to any open-source data collection effort on cyber conflict patterns in the form of lagged information about cyber threats that occur in clandestine settings, such an effort would regardless lead to a useful resource useful to cyber-security scholars across a range of disciplines, upon which others can build in the future.

3. THE UTILITY OF OPEN SOURCE DATA COLLECTION

Open-source collection of information on cyber conflict processes represents the future of data generation in the field, but also presents many challenges. Whereas most open source data collection seeks to parse signal from noise, a cyber conflict effort will miss things simply because not all of the signals are observable from the public sphere.

Why should researchers even attempt to undertake open-source collection of information on cyber conflict trends, given the inherent problems in doing so? We argue that there are three reasons. First, social science research on cyber conflict requires a foundation of knowledge from which to build and infer. Second, assessing open-access description of cyber conflict allows researchers to look at both the content and context of cyber interactions. Third, there are distinct benefits to a scaled-up approach to studying cyber conflict over traditional small-n approaches, as there is additional clarity and opportunity to use advanced analytic methods to parse observable relationships.

The Need for a Knowledge Foundation

Most simply, there is a clear need for foundational knowledge about cyber conflict. At present, there is a relative lack of empirical work in the domain that presents a comprehensive and systematic description of the global impact of the information revolution. One clear argument in favor of scholarly attempts to build a representation of such processes via collection of public-facing information is quite simply that scholars are duty-bound to utilize any resource available in trying to contribute to the condition of knowledge on a given topic.

More pressing than the duty of social scientists, however, is the need to develop knowledge foundations in order to spur the development of a robust research

program. The nature of the development of research programs is a source of hot debate among both classical and current philosophers of the social science enterprise. It is generally acknowledged, however, that research programs are layered bases of theoretical knowledge where peripheral hypotheses linked to core suppositions are appraised with the aim of advancing the state of a given field (Jackson, 2008). Often, hypothesis testing results in rapid rethinking of specific assumptions such that there is a revolution in macro knowledge. In the debate about progress in the field of International Relations, Lakatos is often invoked as the exemplar for establishing which theoretical ideas are of value over others (Vasquez, 1997). This view requires the development of a theoretical and empirical core, which then is investigated with the purpose of seeking advances over prior investigations. Advances can be examined in the context of providing more theoretical and empirical context over past efforts (Lakatos, 1970).

At present, the research program on cyber conflict is still in its infancy. The condition of general core knowledge at the heart of the research program is remarkably unclear, which suggests that there is a strong imperative to articulate macro-theoretical perspectives. Given this, the need for projects that aim for comprehensive modeling of the scope of global cyber conflict is particularly pronounced.

The Context of Cyber Conflict

Building from the perspective that meaning emerges from the interaction of empirical dynamics and the human treatment thereof, researchers should attempt to undertake open-source collection of cyber conflict trends. Such an approach will inevitably capture more than just the actuarial detail of cyber incidents offered by thick case descriptions; specifically, open-source data collection allows researchers the opportunity to understand the context and content of cyber conflict dynamics more fully. Via the capture of textual metadata, cataloguing of adjacent conflict phenomena, and more (Hopkins and King, 2007), open source data modeling of cyber conflict trends (given relevant controls for duplication of information) offers the ability to understand the nature of information about cyber conflict that exists in the public sphere. Social science scholars of cyber conflict are, for instance, naturally interested in how framing of conflict influences the discourse and deliberation of policymakers, practitioners, and the general public. Is a particular cyber event over-reported in news media? What kinds of information are used in public discourse to construct attribution cases, and do these assessments vary given the context of, say, ongoing foreign policy spats with particular foreign countries? Do certain kinds of attacks receive more negative coverage, and how are relevant stakeholders discussed in such coverage? Understanding such dynamics is critical to efforts that aim to gain a systematic understanding of public reactions to cyber threats, the manner in which the citizenry ascribes responsibility for cyber security to public or private sector actors,

and more. Public-facing data promises an ability to answer fundamental questions about the relationship between cyber conflict and the sociopolitical environment in which foreign policymaking and strategy development take place. Answering such questions should be of paramount importance to scholars.

The Benefits of Scale

Finally, efforts to scale up data collection using computer coding, web scrapping, and machine learning exponentially increase the available data. This universe of big data provides an empirical foundation from which to sort signal from noise in a way that is difficult to do where less input data is involved. This effort requires narrowing search terms based on automated construction of parameters and machine learning, followed by subsequent Bayesian updating of the process based on human review and validation of subsets of input data (as described in Hopkins and King, 2010; Ward, Beger, Cutler, Dickinson, Dorff, and Radford, 2013). At the level of the research project, the benefits of such an approach are obvious. With ICEWS, researchers reported a 50% increase in accuracy with semi-supervised approaches using large amounts of input information over those that had previously attempted only to have machines sort raw data. In essence, sophisticated application of an ontological understanding of conflict processes in coding massive amounts of data allows dissection of information in a way that is not possible with small samples.

At the level of the research program on cyber conflict processes itself, the clear benefit of scale is clarity. Given inherent attribution issues associated with cyber incidents, researchers need to cast their net as wide as possible to include not just major media outlets, but also government documents and cyber security reporting. Cyber security firms in particular are a critical source of reporting. These third-party firms have a financial and reputational incentive to report on the nefarious acts of government operatives online. They are constantly monitoring and looking to expose major intrusions (see, for instance, Kaspersky, 2015). Shifting to a machine-coding scheme that collects disparate sources brings these perspectives together in building a cyber security incident database. The combined observations, even if still imperfect, are orders of magnitude better than any one reporting line.

Put together, each argument for the construction of a larger-event based dataset of cyber interactions is not only needed, but prudent and responsible. The production of knowledge is a process fraught with friction, but we can reduce the hindrances common at the start of such enterprises by seeking to establish an empirical baseline early in the lifespan of a research program. Now we move to a formal description of how such a process of data collection takes place, and observe our results in the pilot study.

4. BUILDING A LARGE-SCALE DATA COLLECTION AND TREATMENT PIPELINE

Machine-coded event datasets such as Phoenix or ICEWS are developed using publicly-available resources.[5] To date, most efforts in political science have used news stories scraped from RSS feeds, repositories like Factiva, and outlet websites. It is, however, possible to draw information from any text resource. Although researchers are likely to favor news stories of various kinds for event data production, it is possible to utilize social media data feeds and information like industry reports.

The production of event data from large corpora is relatively straightforward. Unstructured information is taken from feeds and repositories using the researcher's favored method of text crawling and fed into a database program. From there, information can be sent in a specified format to a program that produces structured, usable event data. A number of such programs exist, but the most well known are TABARI/PETRARCH/PETRARCH2, a series of Python-based programs that treat text and produce data. The function of these programs is also relatively straightforward. Text inputs are broken down to the level of individual sentences and are parsed to produce an XML input that includes both the original text and a language element breakdown. From there, files are passed through the main program, which references a series of preset dictionaries to produce structured data. The dictionary inputs represent the expected vocabulary pertaining to a given topic and are designed by the researcher.[6] The resultant structured data are then usable by researchers or are available for further enrichment. Up to the point described here, data output by a program like TABARI would include event description, source and target information, and metadata (date, source of information, etc.). Further enrichment of this data for the purposes of understanding the context or surrounding content can then be achieved via further application of a range of text modeling, entity extraction, and topic modeling tools, with human interaction only required when specifying input text or when making a particular effort to enrich descriptive event data.

5 The same is true for both data based on the Conflict and Mediation Event Observations (CAMEO) framework (Gerner, Schrodt, Yilmaz, Abu-Jabr, 2002) and the Global Database of Events, Language, and Tone (GDELT) (Leetaru and Schrodt, 2013). These efforts, and earlier ones like the Conflict and Peace Data Bank (COPDAB) and the World/Event Interaction Survey (WEIS) (Azar, 1980), provide granular information on human behavior drawn from an immense collection of available public sources of input data. CAMEO and other frameworks are employed for the purposes of structuring and making sense of the resultant information for analytic purposes.

6 Recently, some advances have been made in automatically generating dictionaries based on the input text (Radford, 2016) specifically in the context of cyber security.

5. THE UNITED STATES' EXPERIENCE WITH INTERSTATE CYBER CONFLICT, 2013-14

In order to demonstrate the utility of such a machine-coded event data production approach to comprehensively scoping the cyber domain, we supplement our arguments here with an application of PETRARCH2[7] to a limited corpus of news stories pertaining to cyberspace and information security issues published in the United States. After discussing our data production effort, we present data below on incidents involving the United States and other countries, and compare our results to those of the only existing cyber conflict data resource (DCID). Though this demonstration is a limited, proof-of-concept effort that focuses on two years and one country's relationships with other countries, we note that results match and arguably outperform those of DCID. Given that this data emerges from a relatively small scrape of available information on national cyber security events, the opportunity for expanded efforts seems clear.

Constructing a Demonstration Dataset Using Machine-Coding

The foundation of our demonstration dataset is a corpus of documents downloaded from LexisNexis. The documents that make up our corpus were selected based on two sets of criteria. First, we select on only United States-based print and wire publications so that we can effectively gauge the viability of a machine coding approach to event data production at the level of an individual country. Second, we collate all news articles that correspond to an extensive formula of keyword collocations that aim to capture all coverage of cyber security issues. The result is an extensive corpus of more than 155,000 news stories across more than thirty years. For purposes of matching outputs to DCID and assessing the viability of a machine-coding approach in the context of the contemporary landscape of cyber conflict, our construction of the demonstration dataset presented below focuses on a two-year period between 2013 and 2014. Specifically, data is drawn from 859,423 input text files at the level of individual statements (sentences).

Raw text taken from LexisNexis is passed through several stages of treatment prior to the output of structured event data. First, text is parsed using the Stanford Core Natural Language Processing (NLP) suite of available programs, which tag named entities and parts of speech (i.e. nouns, adjectives, verbs, etc.) found in the text (Manning, Surdeanu, Bauer, Finkel, Bethard, and McClosky, 2014). The parsing process outputs an XML file that details a breakdown of different language elements. This provides the constituency tree parse necessary for event coding using PETRARCH2. Then, a glue program is used to format raw text chunks and the parsed language information into a file format specified by the authors of PETRARCH2 (see *inter alia* Beieler, 2016). Finally, these files are passed to PETRARCH2 for analysis. Analysis of text fragments at the level of sentences works via reference to a series of dictionaries to which the

7 See https://github.com/openeventdata/petrarch2.

program refers. These dictionaries contain vocabulary for types of conflict actions to be coded, agent types to be considered, and actors that might specifically be identified; the dictionaries can be automatically generated (Schrodt, Beieler, and Idris, 2014; Radford, 2016) but are generally updated manually by the researchers, as was the case here. The resultant data output includes information on the type of conflict action recorded, the source of that action, the target of that action and metadata pertaining to the incident (date, type of agent in the context of a particular actor, etc.).

Resultant Data on U.S. Experience with Interstate Cyber Conflict, 2013-14

Our demonstration set of incident records includes 512 distinct events for the two-year period between January 1, 2013 and December 31, 2014. Of those events, 279 events pertain directly to the United States insofar as the machine-coding process identifies either the originator or target as being American. This is not to say that the United States government or a particular federal entity is linked with every event; rather, this number refers to any actor (often named but sometimes an unknown hacker) that is identified as having a relationship with the United States (i.e. an American firm, individual or domestic person, for instance). Of events that link an incident directly to the United States (as a discrete entity) or the U.S. government, the U.S. is coded as the originator of a cyber conflict incident in 151 instances, and as the target in 91 instances.

FIGURE 1. NUMBER OF CYBER CONFLICT INCIDENTS INVOLVING THE U.S. (TOTAL), 2013-14.

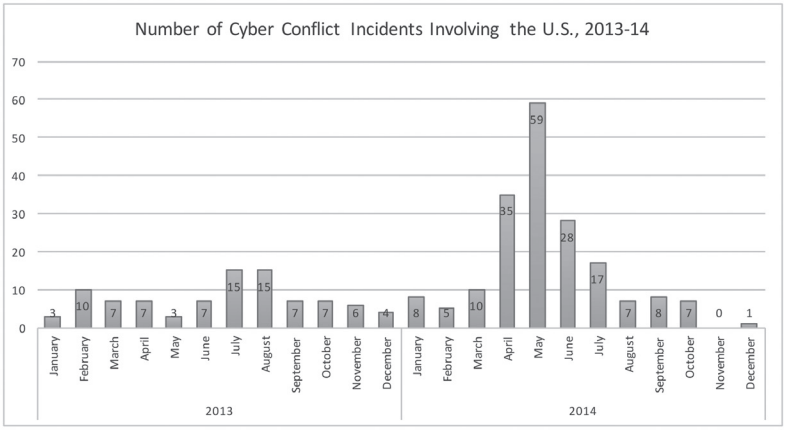
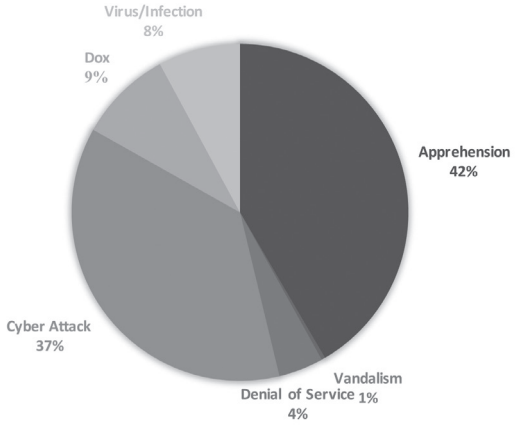


Figure 1 presents the raw count of incidents involving the United States (total attribution, not only government or national attribution) captured in our demonstration machine-coding effort for the years 2013 and 2014. Of these 279, the bulk are identified from March through July of 2014. This is perhaps unsurprising, as this constitutes

the period of time immediately following data breaches at Target, Inc. The Target hacking episode stands as one of the first major instances of a major private firm in the United States going public with the theft of information pertaining to millions of consumers. This period also follows the release of information by Edward Snowden at the end of 2013 pertaining to U.S. cyber operations and electronic surveillance programs, as well as intrusions at the Office of Personnel and Management (OPM) which would stay secret until early 2015. It is worth noting, however, that this data includes both government and non-government activity as captured in open-source reporting, potentially including criminal actions and espionage.

FIGURE 2. TYPES OF CYBER CONFLICT EVENTS CAPTURED INVOLVING THE UNITED STATES, 2013-14.

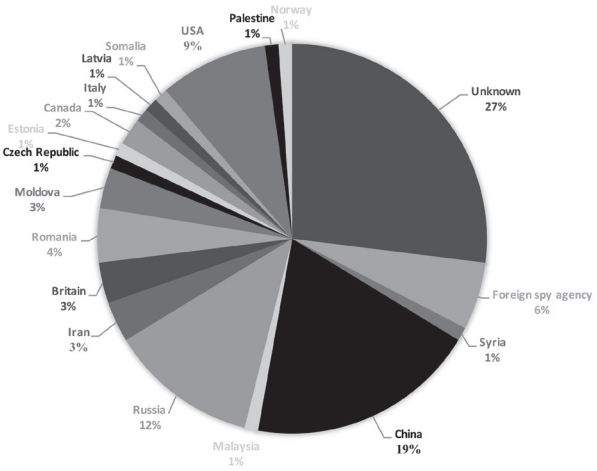


Our test dataset also captures information about the nature of different conflict actions. At the highest level, our approach presents the researcher with six categories of cyber actions – denial of service, vandalism, generic cyber intrusion, malware usage/infection, information doxing, and the apprehension of an involved actor. The denial of service and vandalism categories capture events that specifically reference the terminology of defacement and DDoS. The infection category captures incidents that reference the discovery or presence of a piece of malware based on a set of preset terms and specific malware instances (added to the program dictionary). Cyber intrusions generically refer to cyber actions linked with terminology indicating use of force (‘attacked,’ ‘hacked,’ ‘breached,’ ‘infiltrated,’ etc.) and can therefore cover a wide array of incident types. Apprehension events include instances where perpetrators of an act are caught, arrested or identified. Doxing events include those wherein information is intentionally leaked or released.

Figure 2 breaks down the set of incidents we found involving U.S. actors (as either originators or targets) in 2013 and 2014. By far, the most common incidents recorded

are the apprehension of actors and generic cyber intrusions. Apprehension incidents are coded in a relatively straightforward fashion in that PETRARCH2 identifies language elements pertaining to the arrest and capture of people. Again, cyber intrusions are coded in such a way that a broad number of methods and techniques can produce a cyber intrusion event (such as hacking, intruding, gaining access, injecting code, etc). By contrast, denial of service attacks and digital vandalism are rare in this data set, whilst the leaking of information and incidence of malware (wherein input text does not suggest an attacking action) are uncommon.

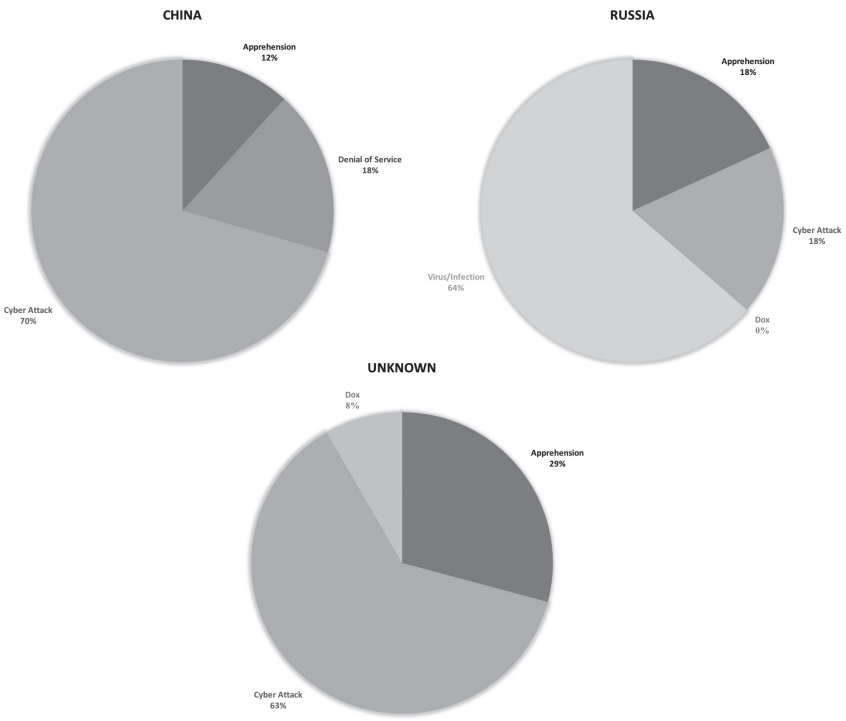
FIGURE 3. SOURCE COUNTRIES FOR CYBER OPERATIONS TARGETING THE UNITED STATES (GOVERNMENT/MILITARY TARGETS).



By means of demonstrating the manner in which machine-coding approaches are useful for capturing attribution within dyads (i.e. where one actor can be seen to have engaged with another), *Figure 3* outlines originator countries for all actions on targets coded as either U.S. government/military targets or 'the United States.' As above, these originator countries are not necessarily identified as government/military/intelligence targets, although many are. It is worth noting that the largest category is 'unknown', where the program is unable to identify a country with which to link a cyber conflict action; this result in itself highlights the attribution challenge faced by researchers in this vein. In almost no instances does this mean that there is no information on the originators of actions; rather, source information is most often tagged at the level of agent types, meaning that no country or specific threat actor can be identified, but the program identifies the originator as a foreign individual or criminal organization. Following this category, the next categories of action are linked with the Russian Federation, the People's Republic of China, and countries linked in analytic work on global cyber conflict with both these countries, such as Moldova and Malaysia. A relatively high percentage of attacks attributed to the U.S. were incidents

where U.S. individuals or groups were involved in cyber conflict actions (mostly being apprehended by authorities).[8]

FIGURE 4. TYPES OF CYBER CONFLICT ACTIONS FOR ATTACKS ON THE UNITED STATES FOR CHINA, RUSSIA AND UNATTRIBUTED INCIDENTS.



Among the three largest originators of conflict actions targeting the United States and its government or military-intelligence apparatus, cyber intrusions are the most common type of event for both China and nationally un-attributable actors. Intrusions might include a wide range of possible techniques, but generally refer to a forceful infiltration without permission, as exemplified in incidents like the OPM hack. With Russia, however, although a substantial percentage of actions linked with the country are generically coded as cyber intrusions undertaken against the U.S., the bulk of coded cyber conflict actions are coded as malware infections. Though such a conclusion is purely speculative, this trend does fit with the narrative of existing research on the nature of global malware distribution, the role of Eurasian organized criminal enterprises in underwriting major ransomware, denial of service and phishing attacks,

8 Regarding the methodological challenges facing the researcher in assessing cyber conflict processes, another point worthy of note off this finding is the degree to which offensive deception is not only possible, but normal. Operators may take steps to mask their point of origin when launching offensive or exploitative actions. See, for instance, http://www.star-telegram.com/news/nation-world/national/article96062667.html.

and on the unique character of the Russian cyber ecosystem that leverages third-party criminal enterprises (Valeriano, Jensen, and Maness, 2017).

Capturing Major Events

The data described above represent only a limited demonstration of how a machine-coding approach to open-source data collection can furnish scholars with unique information about the scope of global cyber conflict. But does the method of approach really function better than traditional human equivalents? Can automated coding of event data match or outperform the research skills of human coders wading through similar information in order to parse signal from noise?

Here, we briefly consider these questions by comparing the results of our demonstration dataset to the preceding DCID cyber conflict data collection effort. Specifically, we ask if incidents involving the United States during 2013 and 2014 that are catalogued in DCID were captured by our initial coding of cyber conflict incidents using an input set of information drawn from all U.S. newspaper sources. Given that our selected input source is news reports, the band of incidents we are most interested in assessing here is those cyber conflict interactions that begin within the period covered (i.e. on or later than January 1, 2013). DCID contains 21 such incidents, which are detailed in *Table 1*.

TABLE 1. CYBER CONFLICT INTERACTIONS (DCID) BEGINNING AFTER JANUARY 1, 2013.

Incident	Start Date	Description
Iron Tiger	1/15/13	Sophisticated APT information theft on US military
Black Coffee	4/1/13	APT17 group hacks Microsoft Tech Net forum
NMCI Hack	9/23/13	Navy and Marine Corps unclassified intranet briefly breached
UConn Hack	9/24/13	Chinese hack steals University of Connecticut data
Operation Pawn Storm	9/30/13	Trojan campaign against Blackwater, State Dep't, and SAIC
Saffron Rose	10/23/13	Information theft campaign on US Aerospace industry
Operation SnowMan	2/1/14	Chinese hackers infiltrate VVW to access military personnel info
Register.com	3/1/14	Register.com, which manages more than 1.4 million websites for businesses world wide, steals network and employee passwords
OPM Hack	3/15/14	OPM hack, personal information of 20 million people stolen
CyberBerkut	3/15/14	Signal NATO members to avoid intervention in Ukraine
Premera Blue Cross	5/5/14	State-sponsored Chinese data breach group steals personal information of 11 million Premera customers
Operation Pawn Storm #1	6/2/14	Backdoor intrusion in to military networks via spear phishing
Operation Pawn Storm #2	6/3/14	Backdoor intrusion in to several commercial networks via spear phishing
US Banks Hack	6/4/14	Retaliation on US targeted sanctions on Russia
UCLA Health Breach	9/1/14	State-sponsored Chinese group, 4.5 million records stolen
White House Hack	10/26/14	White House email server compromised,
DHS Hack	11/6/14	25,000 DHS employees' information stolen from OPM
USPS breach	11/8/14	Personal information of 800,000 USPS employees compromised
State Dep't hack	11/15/14	State Dep't unclassified email system breached and contained
Sony Hack	11/24/14	Sony Pictures is breached and secretive information leaked
Anthem Breach	12/10/14	Black Vine hacker group (China-sponsored) steals sensitive information from health insurance giant Anthem

Our demonstration dataset produced and presented here records events pertaining to 14 of the 21 cyber conflict interactions beginning after January 1, 2013 in the DCID dataset (see *Table 2*). Importantly, incidents not captured by the machine-coding treatment of news stories from the United States largely fall at the end of the period covered. This implies that non-capture is the result of a delay in reporting cyber incidents, and that this issue will be alleviated by a larger time span examining disclosures that happen at a later date (as with the OPM hack, which was revealed in 2015). Moreover, the demonstration dataset contains 1.301 events for each interaction described in DCID, meaning that the average incident described there is matched by more than one reported interaction (even after controlling for duplicates) in the machine-coded version. For instance, the University of Connecticut hack in 2013 was caught twice, with one event annotation describing the infection of computers at the institution, and a later report describing a purposive cyber intrusion aimed at stealing user information.

TABLE 2. CYBER CONFLICT INTERACTIONS IN DCID
(BEGINNING AFTER JANUARY 1, 2013) CAPTURED BY DEMONSTRATION SET.

Incident	Start Date	Description	Recorded?
Iron Tiger	1/15/13	Sophisticated APT information theft on US military	Y
Black Coffee	4/1/13	APT17 group hacks Microsoft Tech Net forum	Y
NMCI Hack	9/23/13	Navy and Marine Corps unclassified intranet briefly breached	N
UConn Hack	9/24/13	Chinese hack steals University of Connecticut data	Y
Operation Pawn Storm	9/30/13	Trojan campaign against Blackwater, State Dep't, and SAIC	Y
Saffron Rose	10/23/13	Information theft campaign on US Aerospace industry	Y
Operation SnowMan	2/1/14	Chinese hackers infiltrate VVW to access military personnel info	Y
Register.com	3/1/14	Register.com, which manages more than 1.4 million websites for businesses world wide, steals network and employee passwords	Y
OPM Hack	3/15/14	OPM hack, personal information of 20 million people stolen	N
CyberBerkut	3/15/14	Signal NATO members to avoid intervention in Ukraine	N
Premera Blue Cross	5/5/14	State-sponsored Chinese data breach group steals personal information of 11 million Premera customers	Y
Operation Pawn Storm #1	6/2/14	Backdoor intrusion in to military networks via spear phishing	N
Operation Pawn Storm #2	6/3/14	Backdoor intrusion in to several commercial networks via spear phishing	N
US Banks Hack	6/4/14	Retaliation on US targeted sanctions on Russia	Y
UCLA Health Breach	9/1/14	State-sponsored Chinese group, 4.5 million records stolen	Y
White House Hack	10/26/14	White House email server compromised,	Y
DHS Hack	11/6/14	25,000 DHS employees' information stolen from OPM	Y
USPS breach	11/8/14	Personal information of 800,000 USPS employees compromised	Y
State Dep't hack	11/15/14	State Dep't unclassified email system breached and contained	Y
Sony Hack	11/24/14	Sony Pictures is breached and secretive information leaked	N
Anthem Breach	12/10/14	Black Vine hacker group (China-sponsored) steals sensitive information from health insurance giant Anthem	N

Given these basic results, we argue that it is reasonable to expect that machine-coding of cyber conflict information can at least match human coder efforts. Indeed, since automated coding of cyber conflict incidents invariably captures the detail of particular actions, it seems reasonable to say that event data production using programs like PETRARCH2 quite clearly outperforms all prior traditional efforts because the scope

is much more comprehensive than a selection on rivals. Specifically, the capture of unique features of different elements of a cyber conflict campaign is a natural byproduct of the heuristic-style approach taken by such programs to describing conflict.

Moreover, machine coding of large quantities of publicly-available and publicly-produced textual information stands to help researchers significantly in addressing attribution challenges with cyber conflict research. Though *political* attribution of cyber attacks is not always feasible and technical attribution is enduringly challenging – if not actually impossible – the use of open source documentation offers researchers advantages on two fronts. First, scale brings with it options for verifying the existence of a particular event (and agency therein) in the form of replicable coding rules that, for instance, only report an incident feature that appears in multiple independent reports. Second, open source data collection generates information that is contextually defined. Regardless of whether or not one considers an effort along these lines to be 100% accurate or not, it is indisputably the case that data collected will reflect the state of public knowledge on a given incident. This is significant because much of what social scientists aim to study with cyber conflict patterns is based on context and perception.

Next Steps

No data collection program of approach is perfect. Both this research team and others attempting to produce a reasonably comprehensive data on global cyber conflict using machine-coding of open source information must grapple with distinct methodological issues over and above the macro challenges of such an approach, as described in the sections above. In addition to this challenge, we must also grapple with the construction of additional independent variables in the composition of cyber security data such as indicators of severity, effects, efficacy, actors, cascades, malware tools, and other associated variables.

From our experience in producing the demonstration dataset employed in this section, we argue that two specific methodological challenges in particular are worthy of attention. First, any major effort to leverage state-of-the-art event data production approaches in this vein must consider the fact that available tools remain relatively dumb. That is to say that tools like TABARI and PETRARCH are entirely focused on extracting meaning from a relatively simple understanding of how language works at the level of the statement. This inevitably leads to errors that need to be checked by human coders when, for instance, the program fails to recognize that a particular event is being offered as a hypothetical.

Correcting such errors might take one of several forms. Simply put, however, the idea for researchers moving forward – the gold standard approach – should be a

hybrid approach consisting of what has been presented here alongside relevant human reliability coding for the purposes of more effectively training algorithms for automated coding. Far from suggesting that researchers use preset understandings of cyber conflict ontologies expressed in dictionaries set by scholarly panels, future work should construct and continually reconstruct the tools of event detection from the collections of information being processed. Doing so will allow researchers to control for several things, not least potential problems with the irrelevance of robustness checks as work is scaled upwards, and the shifting terminology – and even the changing nature – of cyber conflict.

Of course, this first challenge leads to additional work for the researcher that might, in the future, be remedied with increased reliance on machine learning augmentations of current approaches. The second (related) major challenge is that researchers aiming to produce event data must recognize that incident capture is often only meaningful alongside the relevant capture of contextual metadata. Enrichment of event data with information about its construction, framing and more stands to benefit researchers from many disciplines and provides deep detail that compensates for the necessary position researchers must take in producing data that will – at least in terms of how much of cyber conflict can truly be observed – be good, but perfect. Moreover, in the research program on cyber conflict, addressing the attribution problem effectively means providing for uncertainty in empirical investigations. Without appropriate efforts to ensure that quality and certitude metrics are provided by researchers alongside a host of metadata on the presentation of raw information pertaining to cyber conflict, efforts to produce comprehensive resources for the research program will be enduringly limited.

6. CONCLUSION

Though the scope and scale of cyber conflict has grown exponentially over the past four decades, scholarly efforts to examine the domain in a comprehensive fashion remain lacking. To some degree, as we have outlined above, this makes sense as there are real challenges for researchers in the form of attribution difficulties, timing of disclosures, and self-interested gatekeepers of useful data. Given these barriers, lack of enthusiasm for and interest in setting up open source efforts to produce cyber conflict event data is understandable.

We have argued, however, that there is both a clear need and a compelling set of reasons for the development of machine-aided, large-scale data production efforts that utilize public-facing information. Though some argue that open source coding of cyber conflict incidents is impossible due to the covert nature of many acts in

the domain, we both argue and demonstrate that this misstates the issue for security researchers. Data coding carried out in this way both (1) parallels the contours of previous data produced on the subject and (2) additionally provides information on the sociopolitical context of cyber operations. In short, not only does the scope of such an approach to data collection promise an ability for researchers to generalize and cross-validate; it also provides the tools to study cyber conflict in its proper international context, examine the tools utilized in each attack, and understand the nested socio-political dynamics at work during cyber conflicts.

Over and above other factors, an effort to provide comprehensive data on the scope of global cyber conflict as it presents in public-facing information sources stands to give researchers the tools needed to build a robust knowledge foundation. At present, the research program on cyberspace and international security lacks an extensive set of core theses and assumptions that can be challenged. Part of the reason that such a core has been slow to develop is that building bridges between otherwise disparate efforts to flesh out specific topics within the research program is extremely difficult without such a comprehensive data foundation. Even if such a foundation were to contain flaws, it would still function as a common platform upon which researchers could situate meaningful research questions and assumptions, contextualize small-n research, and critique methodological approaches. Naturally, this kind of methodological approach will not include – but rather will stand to augment understanding of – the 'thick' context of cyber conflict, from strategic and institutional cultures to cognitive processes. As projects from Correlates of War to those of the Political Instability Task Force have demonstrated on numerous fronts, however, event data and inferences made from them are necessary elements of field-defining research.

Finally, such an effort to build open source data resources also directly stands to benefit policymakers and practitioners. In addition to the clear added value that comes with improved scholarly knowledge of a given topic, academic data resources might be used by both public and private sector actors as a reference to help excise conjecture from the discourse. An academic basis of knowledge on cyber conflict, founded on a common data resource, affords practitioners the opportunity to involve themselves in scholarly and public debate on issues that can be corroborated without surrendering private information advantages.

7. REFERENCES

"Equation Group: The Crown Creator of Cyber-Espionage," Kaspersky Labs, (February 16, 2015).

Azar, Edward E. "The conflict and peace data bank (COPDAB) project." *Journal of Conflict Resolution* 24, no. 1 (1980): 143-152.

Beieler, John. “Generating politically-relevant event data.” *arXiv preprint arXiv:1609.06239* (2016).

Boschee, Elizabeth, Jennifer Lautenschlager, Sean O’Brien, Steve Shellman, James Starz, and Michael Ward, “ICEWS Coded Event Data”, doi:10.7910/DVN/28075, Harvard Dataverse, (2015), V22.

Brodsky, A. E. “Negative case analysis.” *The SAGE encyclopedia of qualitative research methods* (2008): 553.

Gartzke, Erik. “The myth of cyberwar: bringing war in cyberspace back down to earth.” *International Security* 38, no. 2 (2013): 41-73.

Gartzke, Erik, and Jon R. Lindsay. “Weaving tangled webs: offense, defense, and deception in cyberspace.” *Security Studies* 24, no. 2 (2015): 316-348.

Gerner, Deborah J., Philip A. Schrodt, Omur Yilmaz, and Rajaa Abu-Jabr. “The creation of CAMEO (Conflict and Mediation Event Observations): An event data framework for a post cold war world.” In *annual meeting of the American Political Science Association*, vol. 29. 2002.

Gompert, David C., and Martin Libicki. “Cyber warfare and Sino-American crisis instability.” *Survival* 56, no. 4 (2014): 7-22.

Hopkins, Daniel, and Gary King. “Extracting systematic social science meaning from text.” *Manuscript available at https://gking.harvard.edu/files/words.pdf* 20, no. 07 (2007).

Hopkins, Daniel J., and Gary King. “A method of automated nonparametric content analysis for social science.” *American Journal of Political Science* 54, no. 1 (2010): 229-247.

Healey, Jason, and Karl Grindal, eds. *A Fierce Domain: Conflict in Cyberspace, 1986 to 2012*. Washington, DC: Cyber Conflict Studies Association, 2013.

Jackson, Patrick Thaddeus. “Foregrounding ontology: dualism, monism, and IR theory.” *Review of International Studies* 34, no. 1 (2008): 129-153.

Kello, Lucas. *The Virtual Weapon and International Order*. Yale University Press, 2017.

King, Gary, 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*, Vol. 57, No. 03, (2002): 617-642.

Kostyuk, Nadiya, and Yuri M. Zhukov. “Invisible Digital Front: Can Cyber Attacks Shape Battlefield Events?” *Journal of Conflict Resolution* (2017).

Lakatos, Imre, “Falsification and the Methodology of Scientific Research Programmes,” in Imre Lakatos and Alan Musgrave, eds., Criticism and the Growth of Knowledge (New York: Cambridge University Press, 1970), pp. 91-197.

Leetaru, Kalev, and Philip A. Schrodt. “GDELT: Global data on events, location, and tone.” In *ISA Annual Convention*. 2013.

Lewis, James Andrew. *Assessing the Risks of Cyber Terrorism, Cyber War and Other Cyber Threats*. Washington, DC: Center for Strategic & International Studies, 2002.

Libicki, Martin C. *Crisis and Escalation in Cyberspace*. Rand Corporation, 2012.

Lindsay, Jon R. “Stuxnet and the limits of cyber warfare.” *Security Studies* 22, no. 3 (2013): 365-404.

Lindsay, Jon R. “Tipping the scales: the attribution problem and the feasibility of deterrence against cyberattack.” *Journal of Cybersecurity* 1, no. 1 (2015): 53-67.

Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. "The Stanford CoreNLP Natural Language Processing Toolkit" In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (2014), pp. 55-60.

Nye Jr, Joseph S. "Deterrence and Dissuasion in Cyberspace." *International Security* 41, no. 3 (2017): 44-71.

Radford, Benjamin James. "Automated Learning of Event Coding Dictionaries for Novel Domains with an Application to Cyberspace." PhD diss., Duke University, 2016.

Rid, Thomas, and Ben Buchanan. "Attributing cyber attacks." *Journal of Strategic Studies* 38, no. 1-2 (2015): 4-37.

Schrodt, Philip A., John Beieler, and Muhammed Idris. "Three's a Charm?: Open Event Data Coding with EL: DIABLO, PETRARCH, and the Open Event Data Alliance." In *ISA Annual Convention*. 2014.

Slayton, Rebecca. "What Is the Cyber Offense-Defense Balance? Conceptions, Causes, and Assessment." *International Security* 41, no. 3 (2017): 72-109.

Stoll, Clifford. "Stalking the wily hacker." *Communications of the ACM* 31, no. 5 (1988): 484-497.

Valeriano, Brandon, and Ryan C. Maness. "The dynamics of cyber conflict between rival antagonists, 2001–11." *Journal of Peace Research* 51, no. 3 (2014): 347-360.

Valeriano, Brandon, and Ryan C. Maness. *Cyber War Versus Cyber Realities: Cyber Conflict in the International System*. Oxford University Press, USA, 2015.

Valeriano, Brandon, Benjamin Jensen and Ryan C. Maness. *Cyber Coercion: The Evolving Character of Cyber Power and Strategy*. Oxford University Press, USA, 2018.

Vasquez, John A. "The realist paradigm and degenerative versus progressive research programs: An appraisal of neotraditional research on Waltz's balancing proposition." *American Political Science Review* 91.4 (1997): 899-912.

Ward, Michael D., Andreas Beger, Josh Cutler, Matt Dickenson, Cassy Dorff, and Ben Radford. "Comparing GDELT and ICEWS event data." *Analysis* 21 (2013): 267-297.

Whyte, Christopher. "Ending cyber coercion: Computer network attack, exploitation and the case of North Korea." *Comparative Strategy* (2016).

Whyte, Christopher, "Out of the Shadows: Subversion and Counterculture in the Digital Age," (PhD diss., George Mason University, 2017).