EXPLAINING THE METHODOLOGY OF THE EXTREMIST CYBERCRIME DATABASE (ECCD)

Steven M. Chermak, Joshua D. Freilich, Thomas J. Holt, Noah Turner, and Emily Greene-Colozzi

This study uses open source, public information to examine nation-state and non-nation-state ideologically motivated cyberattacks performed against US targets from 1998 to 2018. We created the Extremist Cyber Crime Database (ECCD) that includes scheme, offender and target codebooks to address gaps in existing research and better inform policymakers. We describe our open source collection procedures, the type of information uncovered, and how we assessed their quality and reliability.

Inclusion Criteria

The following criteria must be satisfied to be included in the ECCD:

First, the attack must have occurred between January 1, 1998 and December 31, 2018.

Second, the attack must have targeted United States infrastructure or target(s). The server targeted must be registered on U.S. soil.

Third, if the perpetrator of the attack is a non-state actor, the attack must have been perpetrated for a specific ideological cause. We are focusing on the following ideologies: far-right extremism[1], jihadism,[2] environmental/animal rights extremism,[3] left wing adherents,[4] and single-issue/secular extremism OR the actor may be state-affiliated; including both formal or informal ties to a national government, intelligence agency, or military unit. The attribution may be made by a cybersecurity company or government agency.

Fourth, at least one of the following attack methods must be used:

Data breach: A loss of sensitive information/Personally Identifiable Information (PII) which can include name, address, SSN, DOB, credit card details, etc., stemming from a hack of data or sensitive systems.

[1] Far-rightists subscribe to aspects of the following ideals: [far-rightists are] fiercely nationalistic (as opposed to universal and international in orientation), anti-global, suspicious of centralized federal authority, reverent of individual liberty (especially their right to own guns, be free of taxes), believe in conspiracy theories that involve a grave threat to national sovereignty and/or personal liberty and a belief that one's personal and/or national 'way of life' is under attack and is either already lost or that the threat is imminent (sometimes such beliefs are amorphous and vague, but for some the threat is from a specific ethnic, racial, or religious group), and a belief in the need to be prepared for an attack either by participating in or supporting the need for paramilitary preparations and training or survivalism. Importantly, the mainstream conservative movement and the mainstream Christian right are not included

[2] Jihadists subscribe to aspects of the following beliefs: Only acceptance of the Islamic faith promotes human dignity and affirms God's authority. They reject the traditional Muslim respect for "People of the Book" (i.e., Christians and Jews). They believe that "Jihad," meaning to struggle in the path of God in the example of the Prophet Muhammad, is a defining belief in Islam and that "lesser Jihad" endorses violence against the "corrupt." Jihadists believe that the Islamic faith is oppressed in both "local and nominally Muslim" governments as well as in non-Islamic nations that occupy indigenous Islamic populations. In addition, the West supports the corruption, oppression, and humiliation of Islam, and exploits the region's resources. They believe that the hedonistic culture of the West (e.g., gay-rights, feminism, sexual permissiveness, alcohol abuse, racism, etc.) has a corrosive effect on Muslim social and religious values. For Jihadists, it is a religious obligation to promote a violent Islamic revolution to combat this assault on Islam by targeting nonbelievers (both Muslims and non-Muslims). They believe that Islamic law, or Sharia law, provides the ideal blueprint for a modern Muslim society and should be implemented in all "Muslim" countries by force. Global jihadists are most concerned with combating the West and the United States in particular, while local jihadists focus on specific regional conflicts.

[3] Eco and animal rights extremists include Individuals or groups that subscribe to aspects of the following ideals: Support for biodiversity and bio-centric equality (i.e., that humans are no greater than any other form of life and have no legitimate claim to dominate earth); the earth and/or animals are in imminent danger; the government and/or parts of society such as corporations are responsible for this danger; this danger will ultimately result in the destruction of the modern environment and/or whole species; the political system is incapable and/or unwilling to fix the crisis by taking actions to preserve American wilderness, protect the environment and support biological diversity; there is a need to defend the environment and/or animals. *NOTE: Environmental rights extremists (primarily) are focused on the environment while animal rights extremists (primarily) are most concerned with the rights of animals*.

[4] Left-wing adherents subscribe to aspects of the following ideals: Marxist and/or Socialist and/or Leninist and/or Stalinist and/or Anarchist beliefs (including individual autonomy and collective equality); support for extreme egalitarianism and/or a classless society and/or workers' and ordinary persons rights; extreme hatred of capitalism and/or corporate malfeasance; extreme hatred of racism and/or a belief that American society in general, and the criminal justice system, especially the police and other law enforcement agencies, in particular are systematically/institutionally racist; an extreme hatred of militarism and/or American imperialism and/or colonialism both abroad and domestically; suspicion of traditional mainstream religions; a belief in Black Separatism/Supremacy and/or militant Black nationalism; support for Puerto Rican Independence, and/or support for changing current American society to alleviate the previous mentioned defects and a belief that revolutionary violence as opposed to participation in the political process is necessary.

DDoS: Distributed Denial of Service attack to knock a resource off-line.

Web Defacement: The changing of the original content of a website to content of the attacker's choosing.

Doxxing: The public release of information about a person(s) to cause harm, embarrass, or annoy.

Other attack methods such as email spamming, hacking of social media, or strobing GIF spamming are included.

The ECCD's inclusion criteria decision tree can be found in **Appendix 1**. This diagram provides a processual illustration of the criteria that we applied to each potential incident as well as how potential cases were filtered in or out of the dataset.

Identifying Incidents

We next identified all cyber-attacks that satisfied our inclusion criteria. We reviewed existing databases, chronologies and listings, official records, law enforcement reports, scholarly works, newspaper accounts/listings, other media's listings, online encyclopedias, blogs, and watch-groups/advocacy reports. We also comprehensively searched the Internet and conducted keyword searches using major search engines like Google, Bing, and Yahoo, and leading newspapers like the *New York Times*, to locate relevant events. In addition, an exhaustive list of cybersecurity and hacker-related reporting portals was developed to identify additional information on these incidents. In total, we reviewed over 120 separate sources to create a listing of all known attacks that satisfied our inclusion criteria.

Searching Incidents and Perpetrators

We treated each scheme,[5] the involved perpetrators, and the targets as a case study with the goal of compiling virtually all public information about the cyber-attack, and the individuals involved. After pre-testing and modifying we created a search protocol with over 80 web-engines encompassing a variety of source types, including cyber-specific outlets, general news media, person-searching websites (e.g. WhitePages), criminal activity sources (e.g. Bureau of Prisons), and cybersecurity blogs. The complete listing of search engines is provided in **Appendix 2**.

Our open source searches uncovered varieties, and at times, substantial amounts of information, though this varied by incident. The information included media accounts; police and other government documents; court records and other materials.

[5] We use the term "scheme" to refer to an illicit cyber operation involving a series of attacks motivated by the same ideological cause or purpose carried out by one or more perpetrators against a target, or a series of targets over a period of time (see Belli, 2012; Freilich, Chermak, Belli, Gruenewald & Parkin, 2014). For instance, a hacker group may compromise a database maintained by a company, then deface the company's website announcing the breach, and then release that data online to all further the same ideological cause. While these are three attacks they are related and thus fall under a single scheme.

Search File Reliability

We implemented steps to enhance the search files quality and reliability. First, we conducted systematic RA searcher trainings to ensure uniformity and reliability across searchers and research sites. Project managers reviewed existing search files to familiarize searchers with each search engine and database. All searchers were taught to properly collect, organize, and store information on each search file. Each searcher was provided a total of 4 – 8 “test cases” to search and told to record all search terms they used. Project managers reviewed and provided feedback on improving their searches, with suggestions such as additional keywords, advanced searching techniques (i.e. date restrictions), and formatting revisions. This process was repeated as many times as needed until the files were sufficient to move forward.

Second, unlike other studies, we include every single piece of information, even tangential and repeat information. Our prior work has demonstrated that as a case investigation and court proceedings progressed, more information became available that resolved contradictory and unclear prior accounts.

Third, we addressed the potential limitation that open-source may include information of varying quality and reliability. Sometimes various source types contain conflicting information and we developed protocols to resolve these inconsistencies. We granted greater weight to the more “trusted” sources following prior that ranked source types by their reliability (e.g., court document versus anonymous blog). Additionally, if two media accounts disagreed, we privileged known outlets, and recognized established local outlets over other media reports following.

Fourth, we created measurement attributes to both enhance the transparency of our search files, and measure each individual file’s reliability. We created a reliability index after we had reviewed many search files and identified which factors characterized those we had more confidence in their accuracy, found in Table 1.

Table 1. Reliability Index for Search Documents

Information	**Presence=1 or Absence=0**
Method of attack was clearly cyber-based	
Court record with factual description of incident	
Government publication/document with factual description	
Perpetrator(s) clearly identified	
Perpetrator profile, background, DOC or related info	
Target(s) clearly identified	
Statement from target(s) about attack	
Target clearly based on U.S. soil	
Total (out of 8)	

Use of this index provided a standardized estimation for the reliability of a search document based on the type of information available, the content of that information, and the relevance of that information to the current study.

Cleaning Process

Upon completion of the search files, we assigned each file to an RA to review the collected documents, “clean them” and prepare the entire file for coding. The cleaning phase served as a buffer between the searching and coding phases, and was used to identify relevant information, conduct additional targeted searches as needed, and make the file more readable for the coder. Students began by reading through the entire search document and adding comments to label pieces of information as indicators of a specific variable in our codebook. For example, if a perpetrator’s name was reported in a news article, the cleaner would highlight that information and add a comment such as “S_2 Suspect Name.” This would be done for every codebook variable that could be identified in the search document. Next, in an effort to be as thorough as possible, students were instructed to conduct follow-up searches to fill in any possible gaps in the information collected. If new information was found in these searches, the student would add it to the search document and repeat the process of identifying and commenting on variable indicators. Finally, the cleaner checked the formatting of the document and ensured all guidelines were satisfied.

Coding

Once the search file was cleaned it was assigned to an RA to review the collected documents and code relevant variables into a flat file. When possible, each data item was triangulated through multiple sources to increase reliability. To help ensure uniformity and accuracy of the coding, coders drew upon a standardized coding instrument and underwent a period of probationary training. Prior to coding, each RA was instructed to review the cleaned search document and repeat the process of identifying variable indicators from the collected information. Once this was completed, the coder would fill out the document reliability scale detailed above and take a count of each source type. Next, the RA would input the suspect and target names or identifiers into a spreadsheet to be assigned identification numbers by the Project Managers. Once suspect and target ID numbers were assigned, the RA began the coding process. Each coder was provided an individual spreadsheet, known as their “Coding Protocol.” This spreadsheet contained all scheme, suspect, and target codebook variables. For each variable, coders had to input a “Coding Decision” which would be the ultimate value(s) coded into the final dataset, as well as a “Coding Justification” where the coders had to provide adequate evidence to support their coding decision. Every single coding decision had to be justified by information provided in the case search document. Once the coders had completed their coding protocol for a case, a Project Manager reviewed each of their coding decisions to provide feedback and revisions. If a coding decision was not adequately justified, the Project Manager instructed the coder to provide more evidence in support of their coding decision. This process was repeated several times during the probationary training period, until the Project Manager approved each coding decision and its respective justification. Once approved, the coder then transferred their coding decisions into the final dataset. These transfers were monitored by the Project Manager to ensure the codes in the dataset matched those in the RAs protocol.

Codebooks

The Scheme codebook includes 53 variables. The information captured in this codebook describes the scheme through a number of characteristics, including the number of entities targeted (either successfully or unsuccessfully), attack method(s) used, ideology underlying the incident, total losses of the hack(s), presence or absence of state-sponsorship, and media reporting patterns.

The Suspect codebook includes 53 attributes that focus on the known characteristics of the perpetrator(s) in the scheme. Any and all perpetrators that were identified as being involved in the scheme were coded for in this codebook, with some incidents having a single suspect and others having over 10. The variables in this codebook related to the demographics of the individual, their ideology, extremist history, and presence on social media and in larger online communities.

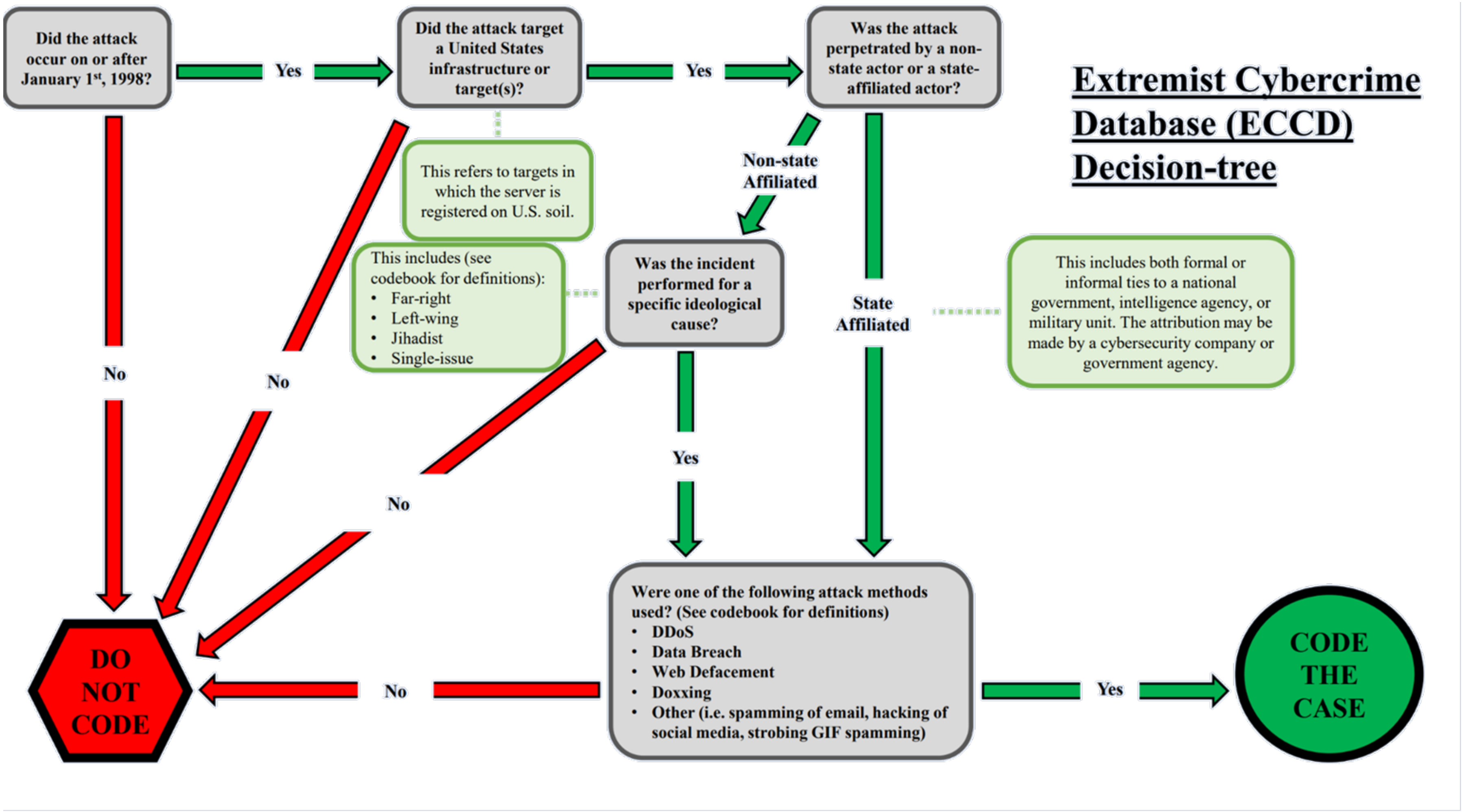
Finally, the Target codebook includes 69 variables. Similar to the suspect codebook, any and all U.S. entities that were either successfully or unsuccessfully targeted in the broader scheme were coded for. Though some schemes involved both U.S. and foreign targets, only U.S. targets were coded for in the Target codebook. Variables related to the type of target (i.e. individual, business, government, educational, military, or transportation), attack type(s) used against them, losses experienced, and characteristics of the targeted server. The codebooks can be provided upon request to the research team.

Funding

This work was supported by the U.S. Department of Homeland Security [ASUB00000368].

Appendix 1.

ECDB Decision Tree



Extremist Cybercrime Database (ECCD) Decision-tree
Did the attack occur on or after January 1st, 1998?
Yes
No
Did the attack target a United States infrastructure or target(s)?
This refers to targets in which the server is registered on U.S. soil.
Yes
No
Was the attack perpetrated by a non-state actor or a state-affiliated actor?
Non-state Affiliated
State Affiliated
This includes both formal or informal ties to a national government, intelligence agency, or military unit. The attribution may be made by a cybersecurity company or government agency.
Was the incident performed for a specific ideological cause?
This includes (see codebook for definitions):
Far-right
Left-wing
Jihadist
Single-issue
Yes
No
Were one of the following attack methods used? (See codebook for definitions)
DDoS
Data Breach
Web Defacement
Doxxing
Other (i.e. spamming of email, hacking of social media, strobing GIF spamming)
No
Yes
DO NOT CODE
CODE THE CASE

Appendix 2.

Search Engines

General Sources	Hacker/Cyber News Sources	Security Vendors & Sourcing	Criminal Activity Sources	Additional Sources	APT/Nation State Specific Info	Blogs
Lexis-Nexis Academic (NEWS & LEGAL)	Hacker news	Trend Micro	State/County DOC Website	Homeland Security Digital Library	MITRE	Krebs on Security
Proquest (AKA Criminal Justice Periodicals)	The Hacker News	Panda media center	Local Police Websites:	Spokeo (person search)	Intel News	Schneier on Security
Google (general)	zdnet	Advisen news	Black Book	Veromi (person search)		
Google News	Threat Post	White Papers	NCSC	Peek You (person search)		
Google Images	Security Magazine	Threat Report	Vinelink	AnyWho		
Google Video	Security News Magazine	Net Scout	Inmate Locater	White Pages		
Yahoo	Dark reading (information IT network)	Github Cyber Monitor	Federal BOP	The 411		
Bing	Tech News World	Github	Mugshots.com	Zaba Search		
Dogpile	Hack Read	PT Security	National Sex Offender Website	Virtual Gumshoe		
Newsbank	The Register	PTAnalytics	BeenVerified	Residential White Pages		
Newspapers.com	Security Week (cyber)	eSecurity Planet	Lexis Advance	Pipl		
News Library	Wired			Pipl (continued)		
newspaperarchive.com	Security Boulevard			Facebook		
WestLaw	Cnet			Twitter		
Courtlistener RECAP	Ciso Mag			Instagram		
USA.gov	Cyber wire			Pinterest		
Police Foundation	Info Security Group			LinkedIn		
PERF	Cyber Security Hub			Blogger		
State and County Courts Websites	Information Week			Wordpress		
Critical Incident and After Action Reports	Cyber Crime Magazine			Technorati (social media portal)		
Google Scholar	Pastebin			National Archives		
				BRBPub		