Identifying the Perpetrator: Attribution of Cyber-attacks based on the Integrated Crisis Early Warning System and the VERIS Community Database

David Noever
Sr. Technical Fellow, PeopleTec, Inc.
Huntsville, AL 35763
david.noever@peopletec.com

David Kinnaird
Computer Science Dept., Univ. Pennsylvania
Philadelphia, PA 19104
kinnaird@sas.upenn.edu

Active cyber-defense is an offensive counter-hack in retaliation for an initial network security breach. In 2015 more than 20 nations declared their plans for enhanced offensive capabilities. This escalation of cyber-weaponry continues despite the inherent internet challenge of plausibility deniability. As described in the dramatically titled "World War C" report from the security firm, FireEye: "The biggest challenge to deterring, defending against, or retaliating for cyber-attacks is the problem of correctly identifying the perpetrator [1]". In Figure 1, the map of nation-state hacking reveals historic and deeply rooted dyadic rivalries translated from the physical to the virtual realm: North-South Korea, India-Pakistan, Iran-Israel, Iran-Saudi Arabia, Russia-Eastern Europe, and US-China. Among nation-states that engage in espionage particularly, the tactic of hacking back depends on a complex event-chain and an equally confounding set of factors: 1) the attacking nation; 2) the victim nation; 3) confidence in source attribution; 4) attack persistence, motive and damage. The present research examines publicly available data to classify attack types, victims and methods.

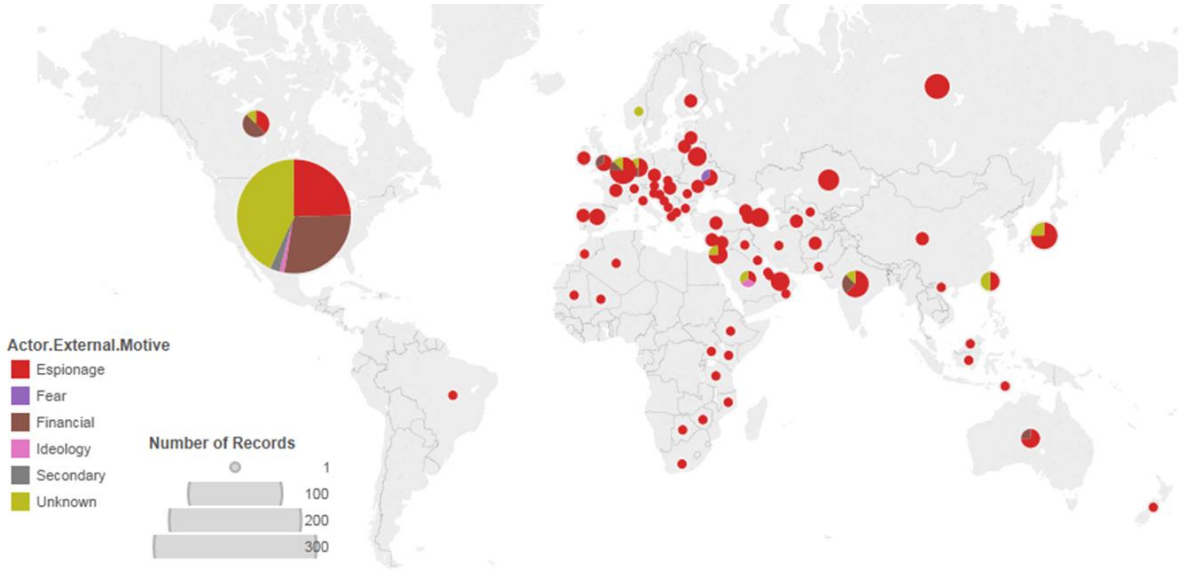

Figure 1. *Geographic distribution of Victim Nations and External Attack Nation Motives from the VCDB.*

From 2000-2014 the VERIS Community Database (VCDB) has reported 5,711 security incidents along with a deep feature set including country attributions for attackers, victims and partners. The VCDB supplies an attributed attack style to at least five actors: 1) nation-states or affiliates; 2) hacktivists; 3) insider threats; 4) terrorists and 5) unknowns. To examine the question of available responses, these community-reported events are compared with news reports that have been machine-classified for sentiment intensities as either positive or negative. The Integrated Crisis Early Warning System (ICEWS) has tracked 14 million news events including 1027 cyber-attacks with targeted country attributes from 1995-2015. By aggregating VCDB and ICEWS incident counts, probabilities can be assigned that a particular tactic originated from a category of cyber-threat actor. Derived from combining the VCDB incident reporting and the ICEWS response frequencies, Figure 2 summarizes the tactics and methods favored by each threat actor type. The most frequent response (36%) is for the victim to make a statement, an ICEWS response scored as neutral. The second most frequent response (18%) is some legal action to arrest or detain the hacker. When no legal action is possible (such as against nation-states), the ICEWS classifications favor the physical rather than virtual responses: 1) investigations (3.2%); 2) unconventional violence (3%); 3) conventional military force (2.2%); 4) deportation (1.2%); 5) rallies (0.8%); 6) diplomacy (0.5%); and 7) threaten with military force (0.5%). For ICEWS data, violence carries the most negative intensity scores. The 31 incidents classified as unconventional violence represent the largest share of nation-state hacking, but notably targeting dissidents not another nation. An initial conclusion from this analysis is that the ICEWS classification system needs extension to a deeper ontology specific to cyberwarfare, akin to the depth found in the VCDB.

To begin building a probabilistic classification for assigning blame based on attack style, a machine learning model was developed based on VCDB incidents. The data was split into training (70%), testing (15%) and validation (15%) subsets. Multiple decision trees [2] were constructed on randomly selected subsets of the training data using the robust classification method called Random Forests [3]. As shown in Table 1, the model output assigns un-normalized, relative weights for the contributing importance of each variable (attack method) to predicted agent to blame (cyber threat actor). Negative importance means the method is inversely probably for a given actor.

Nation-State/ State-Affiliate

- Asset = Server, Network, Interruption
- Vector = Backdoor,Malware, C2, Crimeware, DoS, Downloader, Phishing, Email, Spyware.Keylogger
- Theft = Credentials, Medical, Payment, System, Confidential Data Disclosure
- Target = Bank, Secrets
- Motive = Espionage, Financial
- Containment Timeline=Months-Years

Terrorist

- Asset = Defacement, Media, Server
- Vector = DoS, Malware
- Theft = Medical, Stored Personal Data, Confidential Data Disclosure, Classified
- Target = Unknown
- Motive = Grudge
- Containment Timeline=Minutes

Insider Threat

- Asset = Media, Kiosk, Server
- Vector = Email, Malware, Credentials
- Theft = Stored Personal Data, Potentially Confidential Data Disclosure
- Target = Executive
- Motive = Misuse,Financial, Bribery, Extortion
- Containment Timeline=Months

Hacktivist

- Asset = Defacement, Interruption, Server
- Vector = Malware, DoS, SQLi,Email
- Theft = Medical, Stored Personal Data, Confidential Data Disclosure, Classified
- Target = Executive
- Motive = Financial, Fun, Grudge
- Containment Timeline=Unknown, Hours to Days to Months

Hack Responses (ICEWS)	Count	Percent	Average Intensity
Grand Total	1027	100.00%	-1.99
Make statement	368	35.83%	0.00
Arrest, detain, or charge with legal action	186	18.11%	-5.00
Accuse	72	7.01%	-2.00
Criticize or denounce	66	6.43%	-2.00
Threaten	39	3.80%	-4.40
Investigate	33	3.21%	-2.00
Use unconventional violence	31	3.02%	-9.00
Make an appeal or request	26	2.53%	3.00
Use conventional military force	23	2.24%	-10.00
Consult	20	1.95%	1.00
Reject	20	1.95%	-4.00
Expel or deport individuals	12	1.17%	-5.00
Demand	11	1.07%	-5.00

Figure 2. *Tactics of a nation-state hack extracted from VCDB and public responses extracted from ICEWS.*

The model's conclusion is consistent with the FireEye report [1]: nation-state hackers are defined by their uniquely sophisticated techniques as they seek network credentials or confidential data. A set of attribution rules can be constructed to guess the level of sophistication, target, information-sought and timelines. For example, nation-states are more likely than other cyber-threat actors to penetrate a government and financial server or network over months to years. The information sought may be credentials, medical records, payment systems and other confidential data. To penetrate an entire network, the attack vector may be not a simple denial of service or website defacement but rather more complex backdoors, malware, phishing emails or spyware key loggers. Instead of disruption, the nation-state's intent is to covertly own the server or network with valid credentials. The lack of complete lifecycle data following breaches—from their detection to initial responses to potential retaliation—is a major limitation to addressing this important question. Future work will correlate responses with specific security breaches so that causal models can supplement these statistical ones.

Actor/ Method	Unknown	Hacktivist	Nation-State	Org Crime	Most Probable	Assign	Least Probable	Disregard
Backdoor	27.38	23.79	45.69	3.13	45.69	Nation-State	3.13	Org Crime
C2	35.96	19.34	41.59	3.10	41.59	Nation-State	3.10	Org Crime
Hacking	-11.67	32.62	44.40	34.65	44.40	Nation-State	-11.67	Unknown
MitM	46.55	4.01	49.44	0.00	49.44	Nation-State	0.00	Org Crime
Email	36.24	20.55	63.18	-19.98	63.18	Nation-State	-19.98	Org Crime
Malware	-3.40	32.53	32.21	38.66	38.66	Org Crime	-3.40	Unknown
Phishing	57.14	18.26	95.17	-70.57	95.17	Nation-State	-70.57	Org Crime
Downloader	27.59	26.76	39.77	5.87	39.77	Nation-State	5.87	Org Crime
SMS	74.49	10.78	44.34	-29.61	74.49	Unknown	-29.61	Org Crime
DoS	-27.37	75.63	28.71	23.03	75.63	Hacktivist	-27.37	Unknown
SQLi	35.71	59.77	7.41	-2.89	59.77	Hacktivist	-2.89	Org Crime
Ransomware	24.34	8.17	8.54	58.94	58.94	Org Crime	8.17	Hacktivist
Software	36.73	7.70	10.09	45.47	45.47	Org Crime	7.70	Hacktivist
Spyware	57.87	33.96	10.05	-1.88	57.87	Unknown	-1.88	Org Crime
XSS	61.22	59.75	-20.96	0.00	61.22	Unknown	-20.96	Nation-State
Pretexting	36.71	8.08	5.69	49.52	49.52	Org Crime	5.69	Nation-State
Website	6.78	0.00	10.76	82.45	82.45	Org Crime	0.00	Hacktivist
Other	64.55	20.12	-8.50	23.83	64.55	Unknown	-8.50	Nation-State
Rootkit	58.61	21.62	3.16	16.61	58.61	Unknown	3.16	Nation-State
Cryptanalysis	259.93	-106.74	0.00	-53.18	259.93	Unknown	-106.74	Hacktivist
Partner	100.00	-86.21	0.00	86.21	100.00	Unknown	-86.21	Hacktivist
Spam	100.18	-25.09	0.00	24.91	100.18	Unknown	-25.09	Hacktivist
Phone	-171.19	-108.47	169.49	210.17	210.17	Org Crime	-171.19	Unknown

Table 1. Probabilistic attack attribution table based on the attack style. The probabilities are assigned using variable importance from a Random Forest model [3] similar to many (500+) decision trees [2] from randomly sampled subsets of the VCDB.

References

[1] Geers, Kenneth, et al. World War C: Understanding nation-state motives behind today's advanced cyber-attacks. Technical report, FireEye, 2014.

[2] Markey, Jeff, and Antonios Atlasis. "Using decision tree analysis for intrusion detection: a how-to guide." SANS Institute (2011).

[3] Liaw, Andy, and Matthew Wiener. "Classification and regression by randomForest." R news 2.3 (2002): 18-22.