A Cybersecurity Dataset Derived from the National Collegiate Penetration Testing Competition

Nuthan Munaiah, Justin Pelletier, Shau-Hsuan Su, S. Jay Yang, Andrew Meneely
Center for Cybersecurity, Rochester Institute of Technology
{nm6061,justin.m.pelletier,ss9382,jay.yang,axmvse}@rit.edu

***Abstract*—Developers, and administrators, can benefit from inculcating an attacker mindset to foreshadow potential security flaws is software systems as they are developed and/or administered. However, the lack of empirical data about real cyberattacks poses a challenge to understanding attacker behavior. In this paper, we describe a dataset captured during the recently held National Collegiate Penetration Testing Competition (CPTC) which can provide some insight into typical attackers' modus operandi. The competition had nine teams competing to compromise an enterprise cyberinfrastructure advertised as belonging to a fictitious ride sharing organization. The dataset contains an export of over 500 million log events (with a compressed size of over 8.4 GB) and 99 virtual machines (with a compressed size of over 135 GB). The dataset is, to the best of our knowledge, the first of its kind providing insights at the application level from the perspectives of both the attacker and the victim.**

***Index Terms*—cybersecurity, security competition, dataset**

I. Introduction

Despite substantial investments in cybersecurity practice and research, the frequency, and severity, of data breaches due to security vulnerabilities in software is on a rise [1]. The years of research in both academia and industry have produced processes, tools, and techniques that assist developers and administrators in improving the security of software systems. However, the attackers, incentivized by lucrative payouts, seem to be innovative in discovering and exploiting critical security flaws in software.

Developers and administrators can benefit from inculcating and applying an attacker mindset (i.e. thinking like an attacker) when developing and/or administering software systems. Speculative reasoning by asking questions such as "How can an attacker exploit a certain feature or configuration setting?" can help foreshadow some of the security flaws, however, the insights from such reasoning may fail to capture the nuances of a real cyberattack by a motivated attacker. If we had access to empirical data to characterize real attacker behavior, we could supplement the speculative reasoning to better foreshadow potential security flaws before they become exploitable vulnerabilities. The challenge, however, is the lack of such empirical data.

Organizations that have fallen victim to data breaches may have access to such empirical data, however, sharing the data with the broader research community may present an unacceptable legal risk to the organization. The empirical data might contain confidential or proprietary details about the organization, or, even worse, evidence of negligence on part of the organization. The data may also contain other sensitive information such as employees' proclivities for clickbait which may have led to the breach or may implicate an employee as being negligent to or complicit in the breach. Irrespective of the reasoning, there is little incentive for organizations to release breach postmortem reports or associated data.

The lack of empirical data about real cyberattacks means that scientific inquiry is relegated to theoretical investigations or is often in risk of confirmation bias as datasets are specifically curated to test a single hypothesis. The cybersecurity community can benefit from the existence of such empirical data to help researchers and practitioners be aware of evolving tactics, techniques, and procedures of attack. As David Bisson stated "... security professionals are realizing the importance of being able to understand the mind of the attacker and what they value in a target" [2].

Our goal in this research is twofold:

1) to curate, and disseminate, a dataset containing empirical data (beyond mere packet captures) that could be used to characterize real cyberattacks, and
2) to leverage this dataset to qualitatively describe each attack as a narrative and to quantitatively express the discoverability of vulnerabilities discovered during an attack.

In this paper, we describe our effort toward accomplishing one part of our research goal—the curation and dissemination of empirical data to help characterize (and ultimately prevent) real cyberattacks.

II. Cybersecurity Competitions

The 2017 Global Information Security Workforce Study [3] predicts that there will likely be a cybersecurity workforce gap of 1.8 million by the year 2022. The need for skillful cybersecurity experts is now more immediate than ever and the cybersecurity competition landscape is one way of helping produce the cybersecurity experts of the future. The benefits of cybersecurity competitions is irrefutable and has been a subject of study in several academic publications. Chapman *et al.* [4] conclude that security competitions can be an effective way to teach and motivate students in offensive skills and mindset. Sommestad and Hallberg [5] discuss the potential for use of cybersecurity competitions as a platform for conducting cybersecurity experiments. Tobey discusses the value of security competitions and a methodology for ensuring the quality of the serious games [6]. Hoffman *et al.* reference a myriad of

logistics issues when designing a competition environment [7], which we have overcome with years of experience administering one of the many cybersecurity competitions.

A. Landscape

As alluded to earlier, there is a lack of dataset containing evidence of real world cyberattacks beyond packet captures. While packet captures may be useful in characterizing the pattern of packet flow during an attack, it provides little information about the attackers' modus operandi. We, as researchers, should go beyond packet captures and gather information that could help us understand attackers' mindset. We chose to leverage the thriving landscape of cybersecurity competitions to gather real world cyberattack information as competition participants attempt to penetrate into (simulated) enterprise cyberinfrastructure.

At a high-level, there are many flavors of cybersecurity competitions some of which are conducted locally, while others are held online. The cybersecurity competitions landscape may broadly be categorized as follows:

- *Defensive*: In these competitions, the emphasis is on participants defending their own network in the presence of active threat from a team of security professionals (the Red Team). Participants are expected to secure their infrastructure, root out persistent threats, and monitor for malicious activity, all while maintaining an operational network. The National Collegiate Cyber Defense Competition (NCCDC[1]) is an example of a defensive competition.
- *Capture the Flag*: The goal in these competitions is for participants to gain access to certain pieces of information (the flags) that have been planted on servers in a way that makes is difficult to access. Participants can use any means necessary to capture the various flags. The High School Capture the Flag (HSCTF[2]) and DEF CON CTF are well-known Capture the Flag competitions.
- *King of the Hill*: King of the Hill competitions typically follow a no-holds-barred strategy where participating teams are expected to defend their own network while actively attacking other participants' networks. The Information Security Talent Search (ISTS[3]) is a classic example of a King of the Hill competition.
- *Challenges*: As the name suggests, the competitions in this category are essentially challenges that are meant to assess the cybersecurity skills of participants. The competitions tend to be online with challenges uploaded periodically. Security Treasure Hunt,[4] U.S. Cyber Challenge (USCC[5]), and National Cyber League (NCL[6]) are some of the premier hosts of challenge competitions.
- *Enterprise*: The enterprise competitions tend to be restricted to cybersecurity professionals held primarily for continued education and training.
- *Penetration*: The penetration category is unique because the participants in penetration competitions play the role of penetration testers tasked to penetrate into simulated enterprise cyberinfrastructure. The participants are not only expected to find flaws but also provide recommendations for mitigating the flaws. The National Collegiate Penetration Testing Competition (CPTC) is a classic example of penetration competitions.

The variety in the cybersecurity competition landscape spans a wide spectrum. Each competition assesses a different skill set and, as a result, provides unique value to the broader cybersecurity community.

While almost all categories of cybersecurity competitions provide an opportunity to gather information about real cyberattacks to some degree, the competitions in the penetration category are best suited because participants in penetration competitions are playing the role of attackers as they attempt to systematically compromise an enterprise cyberinfrastructure using any and all available attack vectors.

III. National Collegiate Penetration Testing Competition (CPTC)

In our work, we chose to target the National Collegiate Penetration Testing Competition (CPTC [8]), in which teams of students develop and hone the skills required to effectively discover, triage, and mitigate security vulnerabilities. The competition provides each participating team a simulated enterprise cyberinfrastructure that is meticulously engineered to be as close to a real world enterprise cyberinfrastructure as possible. The competition infrastructure is designed to include services and applications that typical enterprises tend to have such as active directory, email, and domain name systems. Some of the applications and services in the infrastructure have security flaws injected into them to allow for participants to penetrate into the infrastructure. We specifically chose to target CPTC in our work because, being the hosts of the competition, we had the ability to instrument the simulated infrastructure to gather the information needed to characterize real world cyberattacks from the perspective of an attacker.

A. Brief History of CPTC

Since the inception of the competition in 2015, the competition has been a collaboration between academicians and professionals in the cybersecurity space. The inaugural year attracted nine teams primarily from the Northeastern region of the United States which included teams from Pennsylvania, Maryland, and New York. CPTC '16 saw ten teams from eight states across continental United States with teams from California to Maine. During the competition, data were collected with the intent of correlating information from security monitoring tools such as NetFlow, Suricata Intrusion Detection System (IDS), and osquery with survey responses from participants and packet traffic captured during the competition.

Owing to the overwhelming interest in the competition, the third iteration of CPTC—CPTC '17—saw the introduction of regional competitions across the Eastern, Central, and

[1] https://www.nationalccdc.org/
[2] https://hsctf.com/
[3] https://ists.ritsec.club/
[4] http://securitytreasurehunt.com/
[5] https://www.uscyberchallenge.org/
[6] https://www.nationalcyberleague.org/

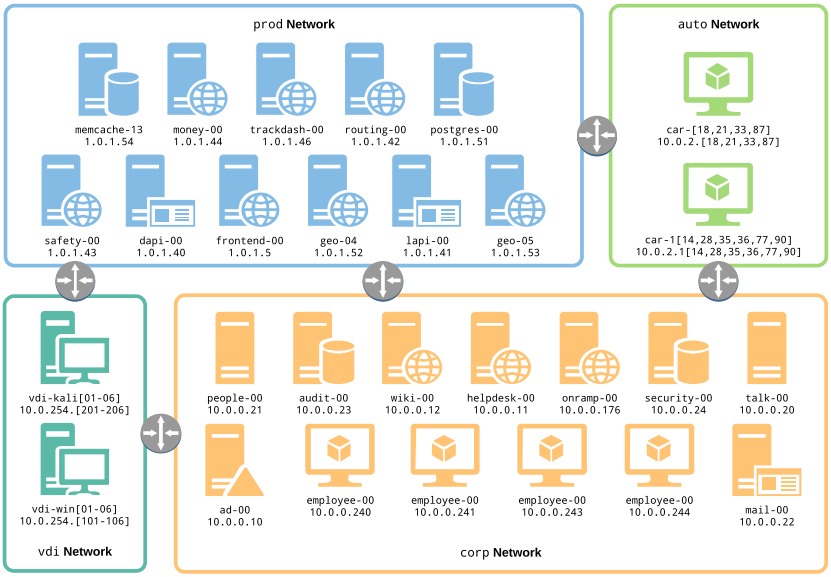

Fig. 1. CPTC '18 network infrastructure that the competition participants had to penetration test

Western regions of continental United States leading up to the national event. CPTC '17 also saw a rise in the quantity of monitoring data collected with over 150k Suricata alerts generated. The data collected during the competition was presented as realtime dashboards to competition judges and members of the advisory board.

B. CPTC '18

As with previous iterations of CPTC, the competition environment was engineered to be as similar to a real world enterprise cyberinfrastructure as possible. The simulated enterprise environment for CPTC '18 was based on a ride sharing company (like Uber or Lyft) with significant research and development in autonomous vehicles. The competition infrastructure was advertised as belonging to a fictitious organization named WHEELZ, the logo for which is shown in Figure 2.

Fig. 2. Logo of the fictitious autonomous vehicle ride sharing organization on which the simulated competition environment for CPTC '18 was based on

The engineering of the competition infrastructure was spearheaded by a volunteer from Uber. The specifics of the security flaws injected into the infrastructure was guided by a select group of volunteers who are professional penetration testers. The security flaws injected were inspired by the kinds of flaws the volunteers have witnessed in their professional engagements. The competition was designed to assess participants' ability to penetrate into the network, application, or social layers of the organization and to write reports and present findings to the executives of the organization. The evaluation spanned the spectrum from initial reconnaissance through to actions on objective stages of the cyber kill chain (applied at each layer). The participating teams that qualified to compete in the Nationals were expected to be prepared to conduct an enterprise-level penetration test.

The network diagram providing an overview of the competition infrastructure is shown in Figure 1. As shown in the network diagram, the competition infrastructure was composed of four subnetworks: corporate (`corp`), production (`prod`), automotive (`auto`), and Virtual Desktop Infrastructure (`vdi`). The participants have complete control over the hosts in the `vdi` network and may use these hosts to penetrate into the other networks.

Each participating team was provided an identical copy of the competition environment with each team isolated (both logically and physically) from other teams. The task of provisioning the competition infrastructure was entirely automated using a custom utility called Laforge [9] developed by one of the competition volunteers. Laforge allows declarative specification of networks and hosts using plaintext files which can then be applied multiple times.

In addition to the tangible network infrastructure, we also manually generated dozens of online personas for members of the fictitious organization. The intention with such an exercise was to include a social aspect to the competition in which participating teams can practice open-source intelligence gathering using publicly-available information. On similar lines, during both the regional and national competitions, we weaved an inject narrative which included asking teams to conduct a preliminary inquiry into allegations of insider threat. The preliminary inquiry was further complimented by seeding pre-

existing exploits on the chat server within the corporate infrastructure network. We also included artifacts of the advanced persistent threat including digital steganography, attributable malware, and a backstory within the e-mail server, all to pique the participants' social investigative curiosity.

IV. Dataset

The dataset being disseminating with this paper was curated by collecting information from the competition environments provided to the teams participating in CPTC '18 Nationals. The information was collected primarily using Splunk, a log aggregator. Each host in the competition environment was instrumented with a Splunk agent which periodically transmitted logs from the host to a central Splunk server for indexing and storage.

The dataset contains two kinds of data: the events indexed by Splunk (exported in JSON format) and the virtual machine export of the hosts in the competition environment (exported as VMDK files). Please refer to Appendices A and B for information on structure of a typical Splunk event and examples of Splunk events from the dataset, respectively.

At the beginning of the competition, we asked participating teams to volunteer to have their data be used for research and six of the nine teams participating in CPTC '18 Nationals consented. Therefore, the dataset being released with this paper only contains the data collected from the six teams that consented to having their data released. In aggregate, the dataset contains over 500 million Splunk events totaling to over 8.4 GB in compressed size and 99 virtual machines totaling to over 135 GB in compressed size. Shown in Table I is the distribution of the size of the dataset by team.

TABLE I
Distribution of the size of CPTC '18 dataset by team

Team	# Events	Size*	# VMs	Size*
Team 1	42,298,868	1.3 (16)	10	14 (18)
Team 2	44,374,576	1.7 (20)	12	17 (22)
Team 5	38,587,468	1.5 (20)	18	24 (34)
Team 7	64,097,668	1.6 (23)	19	27 (36)
Team 8	241,479,383	1.1 (17)	20	27 (38)
Team 9	95,108,425	1.2 (23)	20	26 (38)
Total	525,946,388	8.4 (119)	99	135 (186)

* Compressed (Uncompressed) in GB

In the interest of privacy, the name of the school is not revealed in the dataset. The teams are referred to by numbers from one through nine that were randomly assigned to each team at the beginning of the Nationals competition. The data collection and dissemination protocol has been reviewed and approved by the Institutional Review Board at Rochester Institute of Technology.

Access to the dataset may be requested by filling out a form accessible at `https://nationalcptc.org/research/`. A link to the dataset will be emailed to the email address you provide in the form.

A. Organization

The dataset is organized in the directory structure shown below:

```
.
├── 2018/
│   ├── tn/
│   │   ├── events/
│   │   │   ├── sourcetype.json.gz
│   │   │   ├── ...
│   │   ├── vmdk/
│   │   │   ├── hostname-flat.vmdk.gz
│   │   │   ├── hostname.vmdk.gz
│   │   │   └── ...
```

The `n` in the name of each directory corresponds to the randomly assigned number associated with a team. The name of the files in the `events` directory identifies the Splunk source type that events in the file are associated with. The Splunk source type in turn identifies the data structure of the event.[7] For instance, `bash_history.json.gz` contains events from the `.bash_history` file. As shown in the dataset directory structure above, every virtual machine exported has two VMDK files associated with. The common subscript of the names of both files identifies the name of the host that the export is associated with. For instance, `geo-05.prod.wheelzapp.com` is an export of the `geo-05` host from the `prod` network.

V. Potential Use Case

One of the key pieces of information collected during CPTC '18 was alerts raised by Suricata Intrusion Detection System which was deployed on all hosts in `corp`, `prod`, and `auto` networks (See Figure 1). Shown in Table II are the statistics of key attributes in the Suricata alerts for each of the six teams that consented to having their data used for research.

Suricata alerts, in particular, or logs, in general, may reflect malicious activities transpiring within an enterprise network. Security analysts typically are overwhelmed by such data due to its volume and noisy nature. More often than not, a significant portion of the data tends to reflect scanning activities, which may or may not be indicative of critical attack behavior. These scans are either not observed or missed by the analysts.

With the alert data, one may pose research questions such as:

- How to extract the subset of intrusion alerts so as to pinpoint the critical attack behaviors?
- How to utilize the extracted attack behaviors for prediction of potential future malicious activities?
- How can the attack behavior analysis inform a more robust network and system configuration?

Examples of research studies using past CPTC intrusion alerts include Moskal *et al.* [10] and Perry *et al.* [11]. These works utilize probabilistic modeling and deep learning approaches to extract, differentiate, and predict attack behaviors. Further research advancements are expected with the additional, complimentary intrusion alert data from CPTC '18. In

[7] https://docs.splunk.com/Splexicon:Sourcetype

TABLE II
STATISTICS OF SURICATA ALERTS GENERATED DURING THE CPTC'18 COMPETITION DISTRIBUTED BY TEAM

Team	Team 1	Team 2	Team 5	Team 7	Team 8	Team 9
# Alerts	40,544	5,777	18,975	10,427	16,465	13,648
# Unique Categories	15	11	11	13	12	12
# Unique Signatures	206	47	136	141	150	148
# Unique Source IPs	29	26	26	25	32	25
# Unique Source Ports	9,952	2,055	1,987	3,279	2,384	3,042
# Unique Destination IPs	43	45	42	46	46	47
# Unique Destination Ports	173	143	102	114	190	293

fact, ongoing research [12] will leverage intrusion alerts along with observations, surveys, and interviews of the intrusion teams to assess and verify the different attack approaches.

VI. SUMMARY

In this paper, we described a large dataset containing over 500 million log events (with a compressed size of over 8.4 GB) captured from six teams that consented to having their data used for research as they participated in the CPTC '18 Nationals competition. The dataset also contains 99 virtual machines (with a compressed size of over 135 GB) exported from the competition environment of the same six teams.

The dataset, to the best of our knowledge, is novel in its size and contents. The information contained in the dataset may be used to characterize, and learn from, the modus operandi of attackers. In understanding the attackers' mindset, developers and administrators can better foreshadow potential security flaws in the software system being developed and/or deployed. We hope that the dataset is useful to the broader research community in proposing novel ways to assist developers and administrators in engineering secure software.

ACKNOWLEDGMENTS

We thank all sponsors of CPTC '18, especially IBM Security, Google, Palo Alto Networks, Eaton Corporation, and 780th MI BDE (U.S. Army). We are also grateful to the volunteers, especially the personnel from Uber, Crowe, Hurricane Labs, NCC Group, IPPSec, Indeed, Iron Net Cybersecurity, University of Buffalo, Xerox, Crowdstrike, and the many faculty, students, and staff members who provided their invaluable support to make CPTC '18 a success.

A part of the data collection and analysis work described in this paper is supported by NSF Awards #1526383 and #1742789 and by the U.S. National Security Agency Science of Security Lablet at North Carolina State University.

The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of NSF, NSA, or any of the CPTC volunteers or sponsors.

REFERENCES

[1] L. A. Gordon, M. P. Loeb, W. Lucyshyn, and L. Zhou, "Externalities and the Magnitude of Cyber Security Underinvestment by Private Sector Firms: A Modification of the Gordon-Loeb Model," *Journal of Information Security*, vol. 2015, no. 6, pp. 24–30, 2015.

[2] D. Bisson, "Should Infosec Professionals Hack To Understand the Mind of the Attacker?" https://www.tripwire.com/state-of-security/off-topic/should-infosec-professionals-hack-to-understand-the-mind-of-the-attacker/, 2015, [Online; Accessed: 2019-01-07].

[3] Center for Cyber Safety and Education, "2017 Global Information Security Workforce Study: Benchmarking Workforce Capacity and Response to Cyber Risk," https://iamcybersafe.org/wp-content/uploads/2017/07/N-America-GISWS-Report.pdf.

[4] P. Chapman, J. Burket, and D. Brumley, "PicoCTF: A Game-Based Computer Security Competition for High School Students," in *2014 USENIX Summit on Gaming, Games, and Gamification in Security Education (3GSE 14)*. San Diego, CA: USENIX Association, 2014. [Online]. Available: https://www.usenix.org/conference/3gse14/summit-program/presentation/chapman

[5] T. Sommestad and J. Hallberg, "Cyber Security Exercises and Competitions as a Platform for Cyber Security Experiments," in *Secure IT Systems*, A. Jøsang and B. Carlsson, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 47–60.

[6] D. H. Tobey, "A Vignette-based Method for Improving Cybersecurity Talent Management Through Cyber Defense Competition Design," in *Proceedings of the 2015 ACM SIGMIS Conference on Computers and People Research*, ser. SIGMIS-CPR '15. New York, NY, USA: ACM, 2015, pp. 31–39. [Online]. Available: http://doi.acm.org.ezproxy.rit.edu/10.1145/2751957.2751963

[7] L. J. Hoffman, T. Rosenberg, R. Dodge, and D. Ragsdale, "Exploring a National Cybersecurity Exercise for Universities," *IEEE Security & Privacy*, vol. 3, no. 5, pp. 27–33, Sep 2005.

[8] Rochester Institute of Technology, "Collegiate Penetration Testing Competition (CPTC)," https://nationalcptc.org/, [Online; Accessed: 2018-11-07].

[9] A. Levinson, "Laforge: Security Competition Infrastructure Automation Framework," https://github.com/gen0cide/laforge, 2018.

[10] S. Moskal, S. J. Yang, and M. E. Kuhl, "Extracting and Evaluating Similar and Unique Cyber Attack Strategies from Intrusion Alerts," in *2018 IEEE International Conference on Intelligence and Security Informatics (ISI)*, Nov 2018, pp. 49–54.

[11] I. Perry, L. Li, C. Sweet, S. J. Yang, and A. Okutan, "Differentiating and Predicting Cyberattack Behaviors using LSTM," in *2018 IEEE Conference on Dependable and Secure Computing (DSC)*, Kaohsiung, Taiwan, (to appear), pp. 10–13.

[12] S. J. Yang, A. Rege, and M. E. Kuhl, "Learning, Simulating and Predicting Adversary Attack Behaviors for Proactive Cyber Defense," *Proceedings of NATO IST Workshop on Intelligent Autonomous Agents for Cyber Defence and Resilience*, vol. 152, pp. 30–34, October 2017, prague, Czech Republic.

APPENDIX

A. Splunk Event Structure

The JSON structure of a typical Splunk event is shown below. At a minimum, all Splunk events in the dataset are expected to have the set of fields shown. The fields whose name begins with an underscore are Splunk internal fields. The non-internal fields shown provide metadata about the event. Despite some of the internal and non-internal fields being irrelevant outside the context of Splunk, we have included them in the dataset for completeness.

```
{
  "_bkt": "",
  "_cd": "",
  "_indextime": "",
  "_raw": "",
  "_serial": "",
  "_si": [],
  "_time": "",
  "host": "",
  "index": "",
  "linecount": "",
  "source": "",
  "sourcetype": "",
  "splunk_server": ""
  ...
}
```

`_raw` and `_time` are the only internal fields that contain information that may be relevant to researchers. The `_raw` field contains the original event data (as a string) and the `_time` field contains the timestamp (as a UTC formatted date and time) of when the event occurred. The non-internal fields provide information about the origin of the event (`host`), the source of the event (`source`), and the format of the event (`sourcetype`). For more information on Splunk default event fields, please refer to the documentation.[8]

In addition to the fields described so far, Splunk automatically parses the original event data in `_raw` for specific recognized `sourcetype` values and extracts a set of predetermined fields. The list of recognized Splunk source types is described in the documentation.[9] If a particular `sourcetype` is not recognized, rules can be defined to specify a way to extract certain fields from `_raw`. For instance, if we know that the `_raw` is in turn a JSON, we can specify a rule to extract fields from within that JSON and make them available as part of the Splunk JSON event structure. These additional fields are shown as `...`.

B. Examples of Splunk Events from the Dataset

Shown below is an excerpt of a Splunk event representing an Suricata Intrusion Detection System alert (`sourcetype` is `suricata:alert`) generated on one of the hosts provided to team 8 (`host` is `t8-corp-talk-00`).

```
{
  "_bkt": "ids~18~B135F5F1-...-723ECD8DFAF7",
  "_cd": "18:42071820",
  "_indextime": "1541271572",
  "_raw": "{\"timestamp\": ... {\"linktype\":1}}",
  "_serial": "18234",
  "_si": ["index01", "ids"],
  "_subsecond": ".440299",
  "_time": "2018-11-03 18:59:32.440 UTC",
  "host": "t8-corp-talk-00",
  "index": "ids",
  "linecount": "1",
  "source": "/var/log/suricata/alert-json.log",
  "sourcetype": "suricata:alert",
  "splunk_server": "index01"
}
```

The JSON event itself does not provide the specifics of the cause of the alert. However, the original event data in the event (in `_raw`) is a JSON formatted string which when appropriately formatted reveals more information as shown.

```
{
  ...
  "_raw": {
    "timestamp": "2018-11-03T18:59:32.440299+0000",
    "flow_id": 1473173274056874,
    "in_iface": "ens4",
    "event_type": "alert",
    "src_ip": "10.0.254.202",
    "src_port": 33830,
    "dest_ip": "10.0.0.20",
    "dest_port": 27017,
    "proto": "TCP",
    "alert": {
      "action": "allowed",
      "gid": 1,
      "signature_id": 2809506,
      "rev": 1,
      "severity": 1
      "signature": "ETPRO ATTACK_RESPONSE MongoDB ...",
      "category": "Successful Administrator Privilege
        Gain",
    },
    "app_proto": "failed",
    "payload": "DgEAAAAAAAA...V0Y2hhdAAA",
    "stream": 1,
    "packet": "QgEKAAAUQg...IV474Exivq",
    "packet_info": {"linktype": 1}
  }
  ...
}
```

Investigating the original event data of the Suricata alert reveals that the purported cause of the alert is "Successful Administrator Privilege Gain" (from `category` field in `alert`). We can also identify the source of the intrusion from the `src_ip` field which, in this case, was 10.0.254.202 (`vdi-kali02` from Figure 1). If we were to go a step further and look at the history of bash commands on `vdi-kali02` after the 2018-11-03 18:59:32.440 UTC (from `_time`), we can see that the attacker attempted to dump a database from the now-compromised MongoDB instance on 10.0.0.20 (`talk-00` from Figure 1).

```
{
  "_bkt": "os~115~309B8BB8-...-EF8B9EC8A780",
  "_cd": "115:119870318",
  "_indextime": "1541273844",
  "_kv": "1",
  "_raw": "mongoexport -h 10.0.0.20:27017 -d rocketchat -o
    rocketdump1 --jsonArray",
  "_serial": "578",
  "_si": ["index02", "os"],
  "_time": "2018-11-03 19:37:24.000 UTC",
  "host": "t8-vdi-kali02",
  "index": "os",
  "linecount": "1",
  "source": "/root/.bash_history",
  "sourcetype": "bash_history",
  "splunk_server": "index02"
}
```

As depicted in these handful of example events, the dataset contains a wealth of information that can help characterize attackers' behavior.

[8] https://splk.it/2BYwSSN [9] https://splk.it/2sbkZVi