Final report for Inaugural Catalyst Fellowship Cohort 2022

Dr. Jeff Schwartzentruber
Sr. Machine Learning Scientist, eSentire
Industry Fellow

CONFIDENTIAL – DO NOT DISTRIBUTE WITHOUT CONSENT

Overview

This report is separated into two main sections, the first discussing my experience during the fellowship program and the second discussing my research contributions. The first section will outline many of the benefits, milestones and accomplishments during the fellowship program, while the second section will present a deeper understanding of the research outcomes.

Section 1: Program Experience

I received many benefits from the program and consider my tenure an overall success. In reflection of the past year there are many attributes of the fellowship that supported my career and research goals, such as network expansions, out-of-domain collaborations and access to additional resources that increased visibility and dissemination of my cyber-security research.

In terms of network expansions, the fellowship fostered strong collaboration among the peers. From an industry perspective, strong foundational research is typically relegated to the academics in favor of faster development cycles that support business objectives. Although industry can typically be more agile, the importance of gaining that deep understanding of a topic is sometimes overlooked. However, when engaged with the fellowship program, those fundamental research conversations are not only expected but encouraged. Having a strong need for myself to continue and contribute to foundational academic research in cyber-security, the fellowship gave me an additional outlet. It was through the conversations of learning and exploration that I was able to find commonalities among my fellow peers, which then generated further ideas and avenues for research collaborations, that will thankfully persist beyond my tenure of the fellowship. In reflection, this network effect is a very important part of the program and one that I imagine will continue to expand and become more impactful in future cohorts. Having the ability to rapidly expand my research network will undoubtedly accelerate cyber-security research and innovation in Canada, and I look forward to my continued participation in this community.

In addition to expanding my research network, I also expanded my research areas-of-interest due to my collaboration with experts that are outside my primary domain. Through my collaboration with the fellows, I was introduced to other aspects of cyber-security; and at a much higher level, compared to my own organic learnings. Having the ability to collaborate with my peers at such an in-depth level not only expanded my avenues of research but enhanced my own research abilities through the introduction of new topics, methods, and perspectives. The fellowship program fosters an open environment where peers can engage in these academic discussions and teachings regarding the state-of-the-art of cyber-security and in-turn further enhance my own research practices via these multi-disciplinary perspectives.

Another amazing aspect of the fellowship program is the ability to access a wide range of resources and groups within the broader Catalyst organization. For example, working with the Catalyst Comms was an amazing experience that resulted in not only three great webinars, but also for myself, a piece being published in the National Post. I can say that this type of content generation would not have been accessible for myself without the support and encouragement of the program. In addition to Marketing and Comms, there are other facets of the Catalyst that were also available such as training and the cyber range. Although in my tenure I was not able to utilize all of them, the ability for that type of support is greatly appreciated. Reflecting on this component of the fellowship, I am in strong favor of increased programing that support the dissemination of the fellows' research and/or resources that support research in general.

In summary, I very much enjoyed my time in the program and consider it a complete success. Given the many benefits of the program (both for myself and mutually), I am excited to see it grow and become an innovation hub for Canadian cyber-security research. I feel fortunate for my time in the program and look forward to further collaboration.

Section 2: Research

This section will discuss the various research aspects of the fellowship. It has been divided into two sections, one regarding collaborative research outcomes and the second discussing my personal research advancements.

Section 2.1: Collaborative Research

During the initial months of the fellowship, much deliberation was given to the collaborative and directive nature for the desired research out-comes of the program. Due to the multi-disciplinary background of the fellows, idealization of topics was somewhat difficult, and this process should perhaps be revised in future cohorts. However, after much discussion, three topics were agreed upon that encompassed the major theme of the cohort, which was academic industry collaboration. The three topics were: i. enhancing academic-industry research partnerships, ii. enhancing cyber-security education training and iii. improving the ability to commercialize cyber-security IP. All topics were very poignant to the objective of the fellowship program, and at some degree, had a perspective that both the industry and academics could contribute.

The first webinar topic was focused on the issues related to academic/industry collaborations and improving knowledge transfer. Of that, my contribution was to expand on the disconnect between ad hoc industry solutions and research, and how could solutions out of research be sped up in their transition to industry. The second webinar, in which I co-lead was focused on the issues surrounding the cyber-education research and the cyber-security talent gap. Lastly, the third webinar was focused on the issues with commercializing cyber-security IP and its challenges. My contribution to that webinar was to develop strategies for successful intellectual property commercialization. In addition to these webinars, several papers have been produced (or in production) for publication to peer-reviewed journals or conferences.

In summary, the research topics ended up being more difficult for myself considering they were out of my primary domain knowledge (which is heavily tech based), but were a good learning experience nonetheless.

Section 2.2: Personal Research

This section will discuss my personal research contributions and developments during my tenure of the catalyst. Most of this research was done early in the fellowship, but as the webinars progressed my research focused was required to shift. However, I have not since picked it back up again and am continuing to work on the project.

Introduction:

A common mantra among cyber-security professionals is - 'the more data the better'. Cyber-security log collection and analysis is a critical function in the modern cyber-security toolset. Having a plethora of data sources allows for better correlation, increased levels of assurance and better investigation quality on security events. However, as the data scales, so does the engineering requirements to ingest and maintain these sources. This process of collecting and organizing security data is called data engineering. The main process behind data engineering is the development of extract-transform-load (ETL) pipelines. These ETL pipelines are typically a combination of configuration files, software scripts and infrastructure provisioning. Typically, the development of one ETL pipeline is required for each data source ingested. As such, it takes a great amount of expertise, infrastructure and costs to build and manage these pipelines. These pipelines are further complicated by the lack of standardization among log sources (both in type and schema), the volume/bandwidth of the streaming data, breaking changes from upstream sources and the method of ingestion. In the end, most enterprises and government agencies are left with very complex, costly and resource intensive solutions to manage and their data, before even having the ability to analyze it. As such, techniques and approaches to reduce this technical debt is an active area of research among all layers of the ETL and OSI model [1].

From a cyber-security perspective, the biggest challenge in deriving insights from log data is the data engineering component and maintaining the data in a consistent and reliable manner. However, once a strong data pipeline has been built, it can generate a massive amount of value via the insights it can provide. Examples include, monitoring the availability of servers, tracking user authentication/authorization, monitoring network activity, etc. Correlating among these sources add further assurance by security teams that the system/business is operating normally and increases the capability of detecting a threat.

Sadly, most organizations never get the chance to fully realize their potential when it comes to log analysis and correlation due to the data engineering difficulties. Larger institutions struggle with the resources to handle their peta-bytes of data, while smaller organizations struggle with the expertise to set up and maintain these systems. Commonly, organizations are left prioritizing on which data sources to ingest. As such, there is a real need for a solution that can simplify ETL pipelines by being un-opinionated on the log type/format, while still deriving valuable cyber-security insights at scale. The current research aims to accomplish this task through the development of a machine-learning system that leverage NLP on log sources and anomaly/outlier detection on streaming data.

***Related previous work*:**

With respect to the traditional approach to security ETL pipeline, the entities/descriptives within a security log are split (parsed) into their perspective fields. Parsing logs is difficult due to the variety of log formats and field types. The parsing process transforms a log from an unstructured/semi-structured format into a structured dataset, on which it can be more readily analyzed. The main advantage of this process is the ability to extract key sources of information and form a very deterministic and opinionated perspective of the log. Due to its *deterministic* nature, the fields can be correlated against existing known security signatures to determine if a threat has occurred. This signature-based method is the oldest and most prominent approach to cyber-security detections and has its many advantages and disadvantages.

With the proliferation of digital services, big data, and machine learning, so came the development of new methods of cyber threat detection. Many of these new detection methods rely on machine learning algorithms, and unlike its predecessor, are *probabilistic* in nature. The properties of this newer approach expands the detection capability of an organization but does not lend itself well to traditional security operations methods (e.g., They must perform root-cause-analysis (RCA) on what caused an anomaly, vs having an explicit indicator as to why a signature was alerted). Due to this probabilistic property, it has many advantages and disadvantages when compared to the traditional method and is now considered a necessary tool in the arsenal of security engineers.

These two approaches are not mutually exclusive and should be viewed as complementary when defining a mature security posture - they are two sides of the same coin. However, both methods are underpinned by this problem of complex data engineering. As such, this has been an active area of research and falls under many synonyms, such as: log-abstraction, syntactic/linguistic log analysis and automate-log-analysis [2]. Compared to other machine learning advancements, the progression in this area of study is not nearly as impressive. Log abstraction typically follows that of traditional NLP research, but due to the underlying nature of the text and use cases (i.e., loglines vs. natural language sentences, and RCA vs. reasoning/translation/summarization, respectively), the research advancements do not transfer well. Additionally, many of the new methods work on large batch data and supervised methods, with a disregard for latency. These disadvantages ultimately invalidate many of these algorithms from entering industry due to their lack of robustness and scalability on streaming data.

Lastly, anomaly and outlier detection are the predominant methods upon which machine learning cyber-security solutions are built [3]. Depending on their implementation, the properties of these methods are well suited for cyber-security, such as the ability to handle high cardinality and their un-supervised approach. Developing methods that link log-abstraction to anomaly detection in a scalable and unsupervised manner requires further research.

The objective of this research is to develop a system that ‘understand’ any log file, extract security relevant information, and apply anomaly detection on these extracted entities. The efficacy of this system will be assessed against public databases and competitions. Future work will include developing visualizations and explainability that aligns itself with traditional security operations.

***Methods and Materials*:**

Developing a system that lends itself well to automated log parsing and anomaly detection, at scale, is complex. The system must have the following requirements:

- Ingest both bulk and streaming data.
- Handle high volume and throughput, with the ability to elastically scale as demand increases.
- Automatically extract security relevant information from the log file and enhanced with enrichments.
- Apply anomaly detection techniques to the extracted and enriched log files

Figure 1 outlines a proposed architecture to satisfy these requirements. The research is currently entering the validation phase, as many of the features and capabilities of the architecture have been implemented.

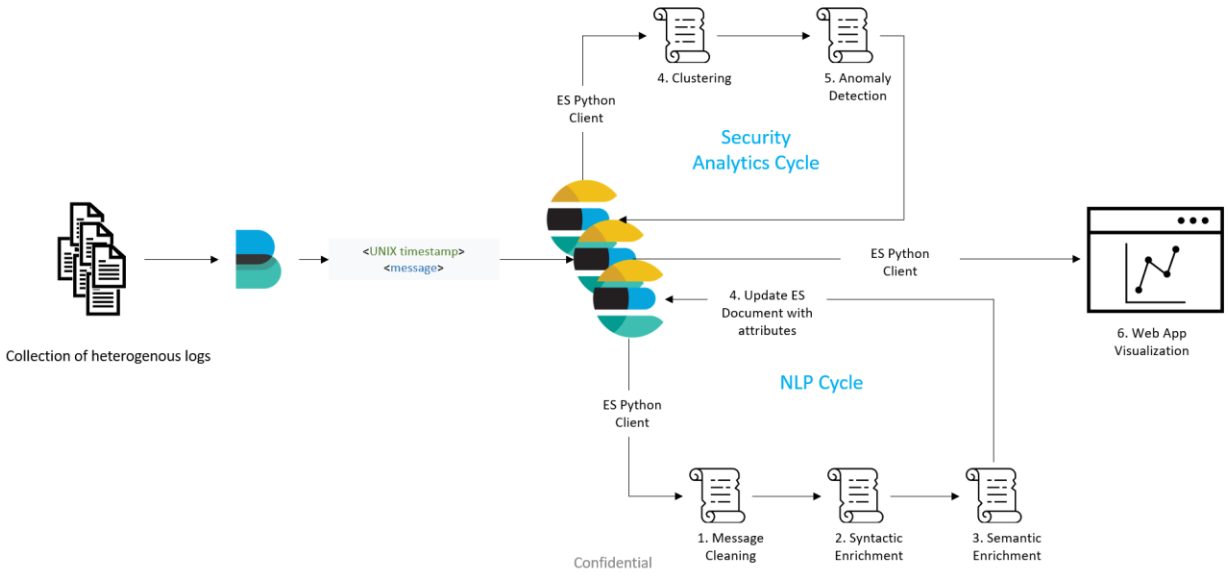

Figure 1. Proposed architecture.

The following list outlines the various functions/capabilities built:

- A single binary agent is configured to ingest the log line as a single, unmodified string for both bulk data files and streaming logs
- A series of regex filters are applied to extract relevant entities.
- roBERT encoding is applied to the log line and last layer extracted.
- Novel clustering algorithm (textStruct) developed to cluster similar log lines (Note: this is a significant addition to the research and will be published on)
- Several Bayesian anomaly detection algorithms applied to the extracted entities.
- Applied in-memory, streaming machine learning algorithms using the RiverML library.
- Containerized and automated the deployment. Setup time on new system < 5 mins.

***Results and Discussion*:**

The following snippet shows a log line, before it has been run through the developed system. It is important to note, that no user-defined logic was applied to the log line – each log would get the same treatment (e.g. unopinionated).

```
1332008993.360000 CD4IXi3Y9PaL1pSk58 192.168.202.138 33771 192.168.27.254 3544
Tunnel::TEREDO Tunnel::DISCOVER\
```

Several calculations (both in terms of formatting, feature engineering and machine learning) are performed on the log line. Due the expansive nature of the system and for the sake of brevity of the paper, the output is shown Appendix A. As you can see, there has been significant information extraction, enrichment and modelling. This increase in information has a significant effect on the quality and fidelity of the anomaly detection capabilities.

For the sake of brevity, only a few of the extract/enriched/modelled results will be discussed. The highlighted blue field (`syntactic.textStruct_wavelet`) represents an array based on the novel *textStruct* algorithm. The intuition behind this algorithm is to exploit characteristics of a log line and allow it to be embedded within a finite Euclidian space, without the need of developing a vocabulary file of the surrounding documents (unlike Word2Vec, BagOfWords, TFIDF, etc.). The specifics of the implementation will be saved for a future paper. As mentioned, the significant contribution of this algorithm is that it is not dependent on the surrounding documents. The vector presented in the syntactic.textStruct_wavelet field was derived based solely on the traditional NLP characteristics of the string itself, and then transformed into the vector space via wavelet analysis. The main advantage of this algorithm is a significant reduction in computational complexity and the ability to work on streaming data, on a per document basis.

Lastly, referring to the highlighted green field (`pla`), this field represents the numeric cluster from three different online streaming algorithms, over three different extracted ‘perspectives/datasets’. Although it has only three values, it is a very involved process, where the ML models results are held in memory and clustering performed on a per log-line (document basis). Additionally, the model hyper-parameters are updated in real-time and adapt to changes in the log data based on the limit of historical data that can be held in memory. The intuition behind these models is to bucket similar log lines within clusters. This helps develop a baseline and identify new and potentially malicious clusters of data within the environment.

Future work will include a new field called ‘anomaly_value’. Although the current system is capable of drawing many conclusions about the anomalousness of a log-line, there is no aggregation of these functions into a unifying term. This value will be an ensemble of many results and is currently being developed. Additionally, leveraging these generated insights to provides better explainability with respect to the traditional security investigation procedure is ideal. Additionally, the results of novel *textStruct* algorithm will be assessed against the LogPAI benchmarks.

***Bibliography*.**

[1] D. El-Masri, F. Petrillo, Y.-G. Guéhéneuc, A. Hamou-Lhadj, and A. Bouziane, “A systematic literature review on automated log abstraction techniques,” *Inf. Softw. Technol.*, vol. 122, p. 106276, 2020.

[2] R. Copstein, J. Schwartzentruber, N. Zincir-Heywood, and M. Heywood, “Log abstraction for information security: Heuristics and reproducibility,” in *Proceedings of the 16th International Conference on Availability, Reliability and Security*, 2021, pp. 1–10.

[3] M. Ahmed, A. N. Mahmood, and J. Hu, “A survey of network anomaly detection techniques,” *J. Netw. Comput. Appl.*, vol. 60, pp. 19–31, 2016.

Appendix A.

The following object outlines the log line after it has been run through the system. The roBERT encoded vector was truncated for brevity.

```
{
  "_index": ".ds-logs-ala-default-2023.01.30-000001",
  "_id": "uksYqIYB7e7YaC03UMv2",
  "_version": 2,
  "_score": 0,
  "_source": {
    "semantic": {
      "roBERTa_msg_raw": [
        0.010528283193707466,
        -0.20581097900867462,
        -0.19983795285224915,
        -0.0920511856675148…
      ],
      "roBERTa_clean_PC4": -0.08059556755886077,
      "roBERTa_clean_PC3": -0.062071769416508014,
      "roBERTa_clean_PC2": -0.09377089730839736,
      "roBERTa_clean_PC1": -0.07211075682521895,
      "roBERTa_msg_clean": [
        -0.0021246301475912333,
        -0.22601978480815887,
        -0.22026042640209198,
        -0.10266334563493729,
        0.14149513840675354…
      ]
    },
    "agent": {
      "name": "elastic-agent",
      "id": "6a471cce-c180-4dfd-a039-4fe3ecef5265",
      "type": "filebeat",
      "ephemeral_id": "a3690350-c24d-459c-aec7-994dee06c4f1",
      "version": "8.4.1"
    },
    "pla": {
      "semantic_cluster_id": 12,
      "syn_sem_cluster_id": 5,
      "syntactic_cluster_id": 16
    },
    "log": {
      "file": {
        "path": "/root/logs/tunnel.log"
      },
      "offset": 30171
    },
    "elastic_agent": {
      "id": "6a471cce-c180-4dfd-a039-4fe3ecef5265",
      "version": "8.4.1",
      "snapshot": false
    },
    "message": "1332008993.360000 CD4IXi3Y9PaL1pSk58 192.168.202.138 33771 192.168.27.254 3544
Tunnel::TEREDO Tunnel::DISCOVER",
    "extracted": {
      "timestamp_isosec": "13"
```

```
},
"tags": [
  "_geoip_database_unavailable_GeoLite2-City.mmdb",
  "_geoip_database_unavailable_GeoLite2-City.mmdb"
],
"input": {
  "type": "log"
},
"syntactic": {
  "token_count": 8,
  "textStruct_wavelet": [
    17.834,
    10,
    9,
    16,
    0,
    0,
    0,
    0
  ],
  "char_count": 110,
  "msg_alphanumeric": 0.12727272727272726,
  "msg_entropy": 4.755142724502403
},
"@timestamp": "2023-03-03T15:30:10.405Z",
"ecs": {
  "version": "8.0.0"
},
"extracted_entities": {
  "ip": "192.168.202.138",
  "loglevel": "ER"
},
"data_stream": {
  "namespace": "default",
  "type": "logs",
  "dataset": "ala"
},
"timestamp_extracted": "13",
"host": {
  "hostname": "elastic-agent",
  "os": {
    "kernel": "5.10.16.3-microsoft-standard-WSL2",
    "codename": "focal",
    "name": "Ubuntu",
    "type": "linux",
    "family": "debian",
    "version": "20.04.4 LTS (Focal Fossa)",
    "platform": "ubuntu"
  },
  "containerized": true,
  "ip": [
    "192.168.16.2"
  ],
  "name": "elastic-agent",
  "mac": [
    "02:42:c0:a8:10:02"
  ],
  "architecture": "x86_64"
},
"event": {
  "dataset": "ala"
```

```
    },
    "pipeline_type": "bulk"
  },
  "fields": {
    "semantic.roBERTa_msg_raw": [
      0.010528283,
      -0.20581098,
      -0.19983795,
      -0.092051186,
    ],
    "syntactic.textStruct_wavelet": [
      17.834,
      10,
      9,
      16,
      0,
      0,
      0,
      0
    ],
    "elastic_agent.version": [
      "8.4.1"
    ],
    "host.hostname": [
      "elastic-agent"
    ],
    "host.mac": [
      "02:42:c0:a8:10:02"
    ],
    "extracted_entities.ip": [
      "192.168.202.138"
    ],
    "host.os.version": [
      "20.04.4 LTS (Focal Fossa)"
    ],
    "extracted.timestamp_isosec": [
      "13"
    ],
    "host.os.name": [
      "Ubuntu"
    ],
    "agent.name": [
      "elastic-agent"
    ],
    "host.name": [
      "elastic-agent"
    ],
    "syntactic.msg_entropy": [
      4.7551427
    ],
    "pipeline_type": [
      "bulk"
    ],
    "host.os.type": [
      "linux"
    ],
    "input.type": [
      "log"
    ],
    "log.offset": [
      30171
```

```
],
"data_stream.type": [
  "logs"
],
"tags": [
  "_geoip_database_unavailable_GeoLite2-City.mmdb",
  "_geoip_database_unavailable_GeoLite2-City.mmdb"
],
"host.architecture": [
  "x86_64"
],
"agent.id": [
  "6a471cce-c180-4dfd-a039-4fe3ecef5265"
],
"ecs.version": [
  "8.0.0"
],
"host.containerized": [
  true
],
"semantic.roBERTa_clean_PC1": [
  -0.07211076
],
"semantic.roBERTa_msg_clean": [
  -0.0021246301,
  -0.22601978,
  -0.22026043,
  -0.102663346,
  0.14149514,
  0.17678666, …

],
"semantic.roBERTa_clean_PC2": [
  -0.0937709
],
"semantic.roBERTa_clean_PC3": [
  -0.06207177
],
"agent.version": [
  "8.4.1"
],
"syntactic.token_count": [
  8
],
"semantic.roBERTa_clean_PC4": [
  -0.08059557
],
"host.os.family": [
  "debian"
],
"syntactic.char_count": [
  110
],
"pla.semantic_cluster_id": [
  12
],
"host.ip": [
  "192.168.16.2"
],
"agent.type": [
  "filebeat"
```

```
    ],
    "host.os.kernel": [
      "5.10.16.3-microsoft-standard-WSL2"
    ],
    "elastic_agent.snapshot": [
      false
    ],
    "elastic_agent.id": [
      "6a471cce-c180-4dfd-a039-4fe3ecef5265"
    ],
    "data_stream.namespace": [
      "default"
    ],
    "pla.syn_sem_cluster_id": [
      5
    ],
    "host.os.codename": [
      "focal"
    ],
    "message": [
      "1332008993.360000 CD4IXi3Y9PaL1pSk58 192.168.202.138 33771 192.168.27.254 3544
Tunnel::TEREDO Tunnel::DISCOVER"
    ],
    "pla.syntactic_cluster_id": [
      16
    ],
    "syntactic.msg_alphanumeric": [
      0.12727273
    ],
    "@timestamp": [
      "2023-03-03T15:30:10.405Z"
    ],
    "host.os.platform": [
      "ubuntu"
    ],
    "extracted_entities.loglevel": [
      "ER"
    ],
    "log.file.path": [
      "/root/logs/tunnel.log"
    ],
    "data_stream.dataset": [
      "ala"
    ],
    "timestamp_extracted": [
      "13"
    ],
    "agent.ephemeral_id": [
      "a3690350-c24d-459c-aec7-994dee06c4f1"
    ],
    "event.dataset": [
      "ala"
    ]
  }
}
```