MIMC: Anomaly Detection in Network Data via Multiple Instances of Micro-Cluster Detection

Rafael Copstein
Faculty of Computer Science
Dalhousie University
Halifax, Canada
rafael.copstein@dal.ca

Bradley Niblett
Executive Consultant
2Keys Corporation
Ottawa, Canada
bniblett@2keys.ca

Andrew Johnston
VP Industry Relations
2Keys Corporation
Ottawa, Canada
ajohnston@2keys.ca

Jeff Schwartzentruber
Faculty of Computer Science
Dalhousie Univeristy
Halifax, Canada
jeffrey.schwartzentruber@gmail.com

Malcolm Heywood
Faculty of Computer Science
Dalhousie University
Halifax, Canada
mheywood@cs.dal.ca

Nur Zincir-Heywood
Faculty of Computer Science
Dalhousie University
Halifax, Canada
zincir@cs.dal.ca

***Abstract*—This paper proposes and explores new attribute correlations and combined effort of multiple instances of micro-cluster-based anomaly detection on port scans, distributed denial of service and botnet attacks. To this end, the proposed system for micro-clustering based anomaly detection is compared against the state-of-the-art technique on three different network datasets, namely CTU-IoT, CTU-13 and UNSW-NB15. Evaluations not only show the effectiveness and high performance of the proposed system on all three datasets but also demonstrate the generalizability of the newly proposed attribute correlations and combination strategies.**

***Index Terms*—Network and service security, anomaly detection, micro-clustering, resilient systems.**

I. Introduction

Network data originated from the capture of logs of online systems is an important source of information for the detection of anomalies. Whether to improve performance, study usage, or to identify possible attack attempts, state-of-the-art (SOTA) techniques have made use of this type of information to establish baseline behaviour in an unsupervised manner.

Previously in [4], we introduced a novel technique, MIMC, and evaluated it against the state of the art (SOTA) technique for the detection of anomalies using network traces. That enabled us to identify the limitations of the SOTA in micro-clustering field and created the opportunity to introduce improvements which were addressed in MIMC. Through empirical analysis, we showed the relevance of the changes proposed by MIMC in detecting anomalous behaviour. Moreover, we demonstrated the ability of MIMC to improve the performance in the majority of cases without sacrificing considerable performance in the others. In [4], we concluded that, in order to better understand the capabilities of micro-clustering, exploring other network-related attribute correlations, as well as the impact of having more than two instances of combination strategies would need further studying.

Therefore, in this paper, we bring two novel contributions:

1) An exploration on the impact in performance when using five new attribute correlations on MIMC over three data sources – including 23 different scenarios – as discussed in Section III-A
2) An analysis on the performance yielded by MIMC when combining more than the two originally proposed instances of micro-cluster detection (an example of this approach is illustrated in Figure 1) – as discussed in Section III-B

The remaining of this paper is organized as follows. In Section II, we introduce the related works from the literature and how they have approached the problem of graph based anomaly detection on network data. In Section III, we analyze the relevance of the new proposed attribute correlations and define our objectives when running the empirical evaluations; in Section IV we present the evaluations and results of the proposed MIMC attribute correlations and parallel instances of MIMC running on sub-graphs; and, finally, in Section V, we draw our conclusions and set a path for future work.

II. Related Works

Detection of suspicious behaviour by analyzing network and service data has been studied over the years by security systems. The surge of data produced by networked systems, services, and Internet of Things (IoT) devices have been growing in all application areas. Most of these data fit into the large and sparse categorical datasets. In the following, we summarize the works from the literature focusing on graph based anomaly detection algorithms to understand these large and sparse data sources.

Noble et al. proposed a graph based anomaly detection technique by detecting recurring substructures in graphs [13]. Mongiovi et al. investigated anomalous regions on dynamic networks [11]. These graphs included traffic networks, social

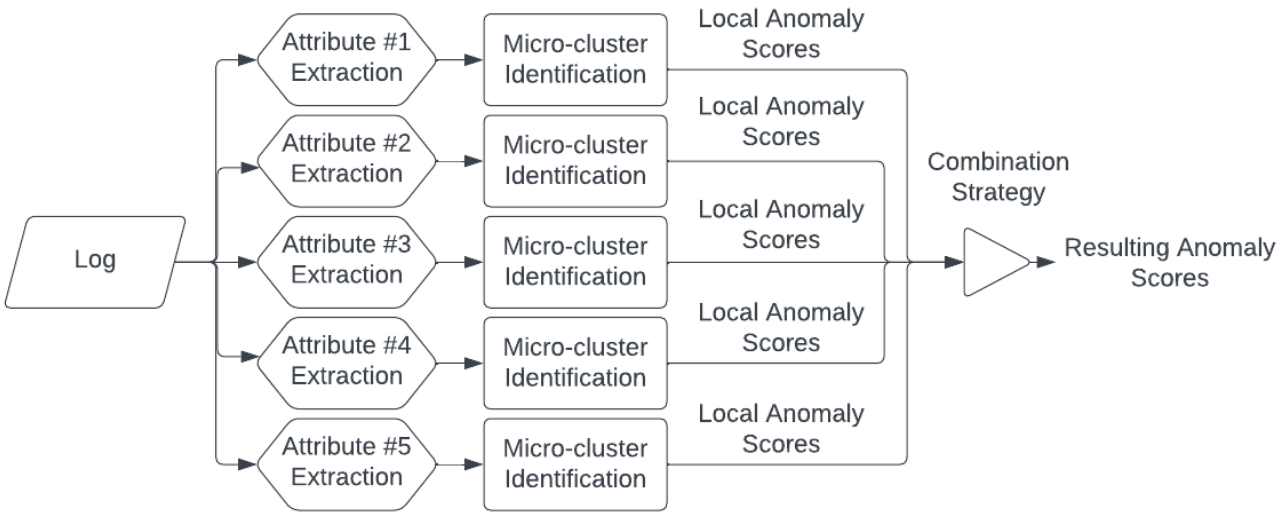

Fig. 1. Overview of the approach used by MIMC to merge local anomaly scores yielded by multiple instances of micro-cluster detection. In the Figure, using five instances

networks, or knowledge networks. Uno et al. explored the idea of micro-clustering, that is, clustering highly related entries as opposed to highly dense ones [16]. They propose a methodology to reduce the number of clusters while maintaining the high relation between data entries. Kulkarni et al. studied the patterns found by creating different kinds of graphs over insider trading data [8]. They used hypergraphs for anomaly detection and showed that the hyperedges identified as anomalies. Lin et al. employed a sensor network data stream anomaly detection method based on optimized clustering [10]. They detected anomalies in the data stream by comparing the information entropy size of the micro cluster and its distribution characteristics. Saebi et al. explored the use of high order networks compared to first order networks for anomaly detection for detecting high order anomalies [14]. Kurniawan et al. employed a knowledge graph to connect knowledge sources and information collected to analyze distributed security logs [9]. Wang et al. developed a model for detecting anomalies using a dynamic micro-clusters scheme [17]. They generated macro-clusters from a network of connected micro-clusters to explore outlier in global and local levels. Bhatia et al. proposed a micro-cluster based anomaly detection as an online method for detecting rapidly arriving groups of similar edges in a dynamic graph [3]. They showed that the detector was good in detecting events such as distributed denial of service attacks in network traffic data.

In summary the research in this field aligns with the overview provided in [1]. They indicate that graph based anomaly detection is effective when data instances are often inter-dependent as well as exhibit long-range correlations, and the anomaly detection problem is often relational in nature, i.e. opportunistic and/or organized crime. Network and service data naturally have these characteristics. Therefore in this paper, we extend our previously proposed graph based algorithm, MIMC, for anomalous event detection [4].

III. Multiple Instances of Micro-Cluster Detection

The state-of-the-art (SOTA) method [3] for detecting anomalies using micro-cluster detection is based on a probabilistic chi-squared test over a dynamic graph of attribute correlations. In other words, for each entry in a data source (logset), a new correlation of attributes is stored on a graph, along with its frequency. If, according to the chi-squared test, the frequency of a given correlation is unlikely, it is categorized as an anomaly.

In this context, an *attribute correlation* is a relation between two co-occurring attributes of an entry in the data source. On the graph, a node represents a specific value found for one of the two attributes in the relation. Each node is connected to all values of the other attribute that have co-occurred with it in entries of the data source. The edge connecting two nodes is non-directional and stores the number of times that same co-occurrence appeared in the dataset.

By making use of a single correlation of attributes per entry in the data source, namely *source IP address* and *destination IP address*, the SOTA method has shown the capability for recognizing a good percentage of anomalies (attacks), especially those where the frequency of packets is suddenly higher than usual, such as distributed denial of service (DDoS) attacks. However, there are other attributes available [2], [7] in network / service data source entries that are not taken into consideration by the SOTA method. Previously [4], we have shown that other attribute correlations – namely *Source Port* $\longrightarrow$ *Destination Port* – do bring positive impact in the performance of micro-cluster anomaly detection. We have also

shown that the combination of two instances of micro-cluster detection with distinct attribute correlations outperforms single instances in the majority of cases.

In the following sections, we show, through an empirical study, the relevance of alternative attribute correlations in detecting anomalies as well as the value in analyzing the results from multiple instances of micro-cluster-based anomaly detection. These two aspects form the foundation of our proposed method **MIMC**.

A. Proposed Attribute Correlations

Some of the most popular attacks directed to application servers target different attributes of the protocol stack. According to the MITRE ATT&CK matrix [15], in order to asses the presence of listening services on a given IP address or IP range, an attacker can perform a port scan by sending multiple packets with varying destination ports on the transport layer and waiting for a response. A successful response indicates the presence of a service listening on that port, while an Internet Control Message Protocol (ICMP) response may be issued in case of the service's absence.

Furthermore, communication from a compromised machine to a botnet Command and Control (C&C) server can be established by making use of non-standard source and destination ports, which may bypass existing firewall rules and filters. Attackers can even leverage third-party services to transmit malicious content to a compromised machine. A malicious software, for example, could be downloaded by making clever use of an Application Programming Interface (API) of a service that allows for user content creation.

Despite the existence of such attacks, the SOTA method only covers the source and destination IP addresses of each packet logged in the data. In order to better understand if the capability of recognizing anomalies on the log data is directly related to this choice of attribute correlation or if other available correlations would be able to harness similar performance, we selected five such correlations that relate to existing attacks in the MITRE ATT&CK matrix. These are:

- Destination IP Address ⟶ Destination Port *(IPDst-PortDst)*
- Destination IP Address ⟶ Bytes Sent *(IPDst-SrcBytes)*
- Source IP Address ⟶ Destination Port *(IPSrc-PortDst)*
- Source IP Address ⟶ Bytes Sent *(IPSrc-SrcBytes)*
- Source Port ⟶ Destination Port *(PortSrc-PortDst)*

The experiments were run over three annotated (with ground truth) data sources containing network traffic flow data of various services and devices:

- **CTU-IoT** [6]: datasets contains 20 malware captures along with benign traffic over IoT devices captured by the CTU University in Czech Republic. Due to the number of flows covered by some of the available captures and timeline constraints, we did not consider the biggest captures 17, 33, 39, and 43 of the dataset, and analyzed the reamining 16 captures (scenarios). The size of each capture varies from tens of thousands to hundreds of millions of packets.

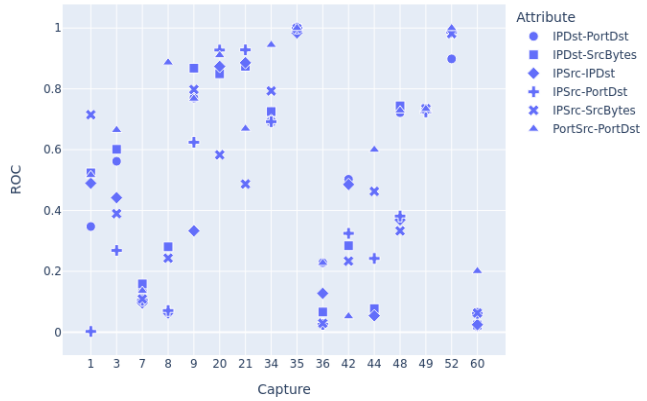

Fig. 2. ROC-AUC values found for each experiment using different attribute correlations over the *CTU-IoT* dataset

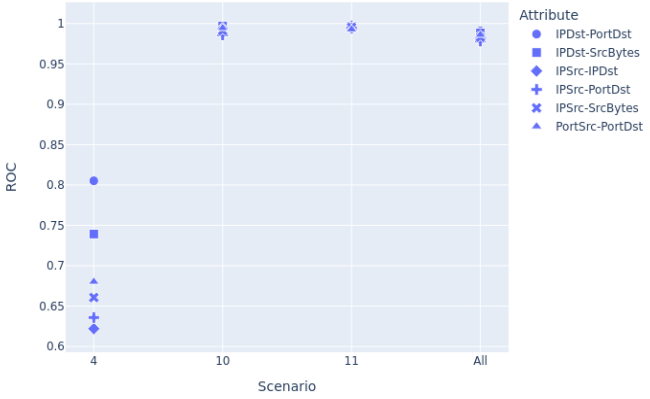

Fig. 3. ROC-AUC values found for each experiment using different attribute correlations over the *CTU-13* dataset

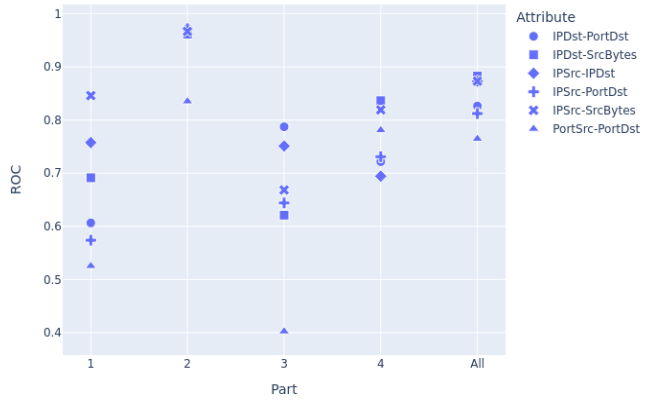

Fig. 4. ROC-AUC values found for each experiment using different attribute correlations over the *UNSW-NB15* dataset

- **CTU-13** [5]: dataset contains botnet traffic captured by the CTU University in Czech Republic. This dataset consists of 13 distinct scenarios of botnet traffic, representing different forms of malicious behaviour. Each of the provided scenarios can be used individually or be combined. SOTA systems considered three scenarios (captures) from this dataset. So to be able to compare the proposed method MIMC to the SOTA, we also considered the same scenarios, namely 4, 10, and 11 of the CTU-13, which are the scenarios containing DDoS attacks. The combined total of the scenarios considered includes $2.5M$ packets that are exchanged between $371K$ hosts.
- **UNSW-NB15** [12]: dataset was captured at the University of New South Wales and contains normal traffic data as well as synthetic modern attack behaviours. This dataset contains approximately $2.5M$ records including, but not limited to, packets for DDoS attacks, backdoor attacks, and fuzzer attacks. This dataset is made available in four scenarios (partitions).

The performance of each experiment was measured using the ROC-AUC calculated over the yielded scores. This allows for better comparison with the original SOTA technique which also presented their results using ROC-AUC measurements.

For the experiments over the *CTU-IoT* dataset – as seen in Figure 2 – there is no single attribute correlation that consistently yields a high performance. Attribute correlations that perform particularly well over one capture do not necessarily perform well for others. In some cases, they might present the worst performance out of the attributes available. The *Source IP Address* $\longrightarrow$ *Destination IP Address* correlation used by SOTA tends to lie in the middle of the results, that is, not being the best available result nor the worst.

For the experiments over the *CTU-13* dataset – as seen in Figure 3 – most attribute correlations perform equally well, with the only exception being Scenario 4, where the best performing correlation is *Destination IP Address* $\longrightarrow$ *Destination Port*, and the correlation used by SOTA ends up with the worst performance of all.

Lastly, for the experiments over the *UNSW-NB15* dataset – as seen in Figure 4 – we, once again, see a good distribution of the attribute correlations in terms of performance over the different parts of the dataset. This time, however, the SOTA attribute correlation appears more consistently as one of the top performers for all parts, whereas the *Source Port* $\longrightarrow$ *Destination Port* correlation more consistently appears as the least performant one.

The objective of these experiments was to show that the use of alternative attribute correlations have merit for identifying anomalous entries in these datasets. That is shown to be the case given that in most of the experiments it is one of the proposed attribute correlations (not the one introduced by the SOTA) that yields the best performance overall. However, it is not the case that a single attribute correlation always yields the best performance. This seems to indicate that we would require their combined efforts in order to improve on the existing performance results.

B. Multiple Instances and Combination Strategies

As seen in the previous experiments, different attribute correlations show better performance on different scenarios of each dataset (data source). In order to achieve consistent high performance, we propose that we merge the scores found by different attribute correlations into a single set of scores aiming to perform better than any of its parts. This proposal is illustrated in Figure 1.

In order to perform the merging of the scores, we also propose three *combination strategies*:

- **MAX**: For the same data entry, keep the highest of the scores.
- **MIN**: For the same data entry, keep the smallest of the scores.
- **AVG**: For the same data entry, calculate the average of the scores.

Each of these is more or less sensitive to differences in scores provided by each of the attribute correlations. For example, if a single attribute correlation yields a high anomaly score, *MAX* will flag that entry of the data source as highly anomalous. On the other hand, *MIN* would require that all attribute correlations evaluated yield a high anomaly score for it to flag the entry as anomalous. *AVG*, in turn, is susceptible to a high standard deviation.

It is worth pointing out that, combination strategy allowing, MIMC is not restricted in terms of the number of scores that can be combined at once. In other words, one can run any number of parallel instances of micro-cluster detection and combine their local anomaly scores into one global score.

IV. Experiments & Evaluation

In order to test the performance of the proposed method, MIMC, we designed an experiment with two main goals:

- Compare the performance of MIMC with that of the SOTA
- Compare the performance obtained by combining attribute correlations with that of the correlations individually

We started by gathering all of the previously proposed attribute correlations into a set and generating all of the possible combinations of this set. Next, we ran MIMC once for each of the combinations of the set where each element of the combination was one of the attribute correlations used in each internal instance of micro-cluster detection of MIMC. For example, given a combination $C = \{\text{IPSrc-IPDst}, \text{IPDst-PortDst}\}$, we ran one instance of micro-cluster detection for the attribute correlation *Source IP Address* $\longrightarrow$ *Destination IP Address* and another for the *Destination IP Address* $\longrightarrow$ *Destination Port* correlation.

Each instance of micro-cluster detection yields a set of anomaly scores related to each entry in the dataset being analyzed. Each of these sets of anomaly scores – called *local* anomaly scores – were then combined into a *global* anomaly score following a *combination strategy*. In this experiment, we tested the three aforementioned combination strategies,

TABLE I
SUMMARY OF RESULTS RUNNING MIMC OVER THE CTU-IOT DATASET WITH THE MAX COMBINATION STRATEGY

Scenario	H.P.C	MIMC Score	SOTA Score	>SOTA	>Individual
1	IPDst-SrcBytes_PortSrc-PortDst	0.5277	0.4896	True	True
3	IPDst-PortDst_PortSrc-PortDst	0.6452	0.4421	True	True
7	IPDst-PortDst_PortSrc-PortDst	0.1356	0.0956	True	True
8	IPSrc-SrcBytes_PortSrc-PortDst	0.9669	0.0629	True	True
9	IPDst-SrcBytes_IPSrc-SrcBytes	0.7973	0.3334	True	False
20	IPSrc-PortDst_IPSrc-SrcBytes	0.9122	0.8737	True	True
21	IPSrc-PortDst_IPSrc-SrcBytes	0.9205	0.8859	True	True
34	IPSrc-SrcBytes_PortSrc-PortDst	0.9501	0.6959	True	True
35	IPDst-PortDst_PortSrc-PortDst	0.9999	0.9839	True	True
36	IPDst-PortDst_PortSrc-PortDst	0.2271	0.1276	True	True
42	IPDst-PortDst_IPSrc-IPDst	0.5017	0.4855	True	True
44	IPSrc-SrcBytes_PortSrc-PortDst	0.6605	0.0539	True	True
48	IPDst-PortDst_PortSrc-PortDst	0.7269	0.3694	True	True
49	IPDst-SrcBytes_PortSrc-PortDst	0.735	0.7261	True	True
52	IPDst-PortDst_PortSrc-PortDst	0.9995	0.9873	True	True
60	IPSrc-SrcBytes_PortSrc-PortDst	0.1902	0.0248	True	True

TABLE II
SUMMARY OF RESULTS RUNNING MIMC OVER THE CTU-IOT DATASET WITH THE MIN COMBINATION STRATEGY

Scenario	H.P.C	MIMC Score	SOTA Score	>SOTA	>Individual
1	IPDst-PortDst_IPSrc-SrcBytes	0.7154	0.4896	True	True
3	IPDst-PortDst_IPDst-SrcBytes	0.6113	0.4421	True	True
7	IPDst-SrcBytes_IPSrc-SrcBytes	0.1587	0.0956	True	True
8	IPDst-SrcBytes_PortSrc-PortDst	0.5679	0.0629	True	False
9	IPDst-PortDst_IPSrc-PortDst_ IPSrc-SrcBytes_PortSrc-PortDst	0.9557	0.3334	True	True
20	IPSrc-IPDst_IPSrc-PortDst	0.9291	0.8737	True	True
21	IPDst-PortDst_IPDst-SrcBytes_ IPSrc-IPDst_IPSrc-PortDst	0.9285	0.8859	True	True
34	IPDst-SrcBytes_IPSrc-SrcBytes	0.793	0.6959	True	True
35	IPDst-PortDst_IPSrc-PortDst	0.9999	0.9839	True	True
36	IPDst-PortDst_IPDst-SrcBytes_IPSrc-IPDst	0.3329	0.1276	True	True
42	IPDst-PortDst_IPDst-SrcBytes_IPSrc-IPDst	0.4867	0.4855	True	True
44	IPSrc-SrcBytes_PortSrc-PortDst	0.5044	0.0539	True	False
48	IPDst-PortDst_IPDst-SrcBytes_IPSrc-PortDst_ IPSrc-SrcBytes_PortSrc-PortDst	0.787	0.3694	True	True
49	IPDst-SrcBytes_IPSrc-IPDst_PortSrc-PortDst	0.737	0.7261	True	True
52	IPDst-SrcBytes_IPSrc-IPDst_PortSrc-PortDst	1	0.9873	True	True
60	IPDst-SrcBytes_IPSrc-IPDst_ IPSrc-SrcBytes_PortSrc-PortDst	0.0698	0.0248	True	False

namely **MAX**, **MIN**, and **AVG**, on each of the aforementioned datasets: **CTU-IoT**, **CTU-13**, and **UNSW-NB15**.

A. CTU-IoT

For the CTU-IoT dataset, we summarized the results into Tables I, II, and III for the *MAX*, *MIN*, and *AVG* combination strategies. In each Table, we observe which scenario (capture) is analyzed, the highest performing combination (H.P.C), the score obtained by MIMC as well as the score obtained by the SOTA method, both calculated as the ROC-AUC of the yielded resulting anomaly scores. Lastly, we observe whether the combination score is higher than the SOTA score or, if not, if it's within 5% of it. In either case the result would be **True**. The same happens for each of the individual attribute correlations that compose the combination. That is to say, we verify that the combination achieves a higher score than any of its elements had when tested by itself.

As we can see from the results, there is only **1** combination in the results for the *MAX* strategy, **3** combinations in the results for the *MIN* strategy, and **1** case in the results for the *AVG* strategy that do not improve over its individual members. In all cases there is a combination that outperforms the SOTA method. It is worth noting that the *Source Port* $\longrightarrow$ *Destination Port* correlation is present in the H.P.C in 12 of the 16 cases using either the *MAX* or the *AVG* merging strategy, indicating that it is a significant contributor to the performance achieved by these experiments.

TABLE III
SUMMARY OF RESULTS RUNNING MIMC OVER THE CTU-IOT DATASET WITH THE AVG COMBINATION STRATEGY

Scenario	H.P.C	MIMC Score	SOTA Score	>SOTA	>Individual
1	IPDst-SrcBytes_PortSrc-PortDst	0.5277	0.4896	True	True
3	IPDst-PortDst_PortSrc-PortDst	0.637	0.4421	True	True
7	IPDst-PortDst_PortSrc-PortDst	0.1356	0.0956	True	True
8	IPDst-SrcBytes_PortSrc-PortDst	0.9436	0.0629	True	True
9	IPDst-SrcBytes_IPSrc-SrcBytes	0.8026	0.3334	True	False
20	IPSrc-PortDst_PortSrc-PortDst	0.9109	0.8737	True	True
21	IPSrc-PortDst_IPSrc-SrcBytes	0.9206	0.8859	True	True
34	IPSrc-SrcBytes_PortSrc-PortDst	0.9484	0.6959	True	True
35	IPDst-PortDst_PortSrc-PortDst	0.9999	0.9839	True	True
36	IPDst-PortDst_PortSrc-PortDst	0.2275	0.1276	True	True
42	IPDst-PortDst_IPSrc-IPDst	0.4972	0.4855	True	True
44	IPSrc-SrcBytes_PortSrc-PortDst	0.6519	0.0539	True	True
48	IPDst-PortDst_PortSrc-PortDst	0.7224	0.3694	True	True
49	IPDst-SrcBytes_IPSrc-IPDst	0.735	0.7261	True	True
52	IPDst-PortDst_PortSrc-PortDst	0.9994	0.9873	True	True
60	IPSrc-SrcBytes_PortSrc-PortDst	0.1882	0.0248	True	True

TABLE IV
SUMMARY OF RESULTS RUNNING MIMC OVER THE CTU-13 DATASET WITH THE MAX COMBINATION STRATEGY

Scenario	H.P.C	MIMC Score	SOTA Score	>SOTA	>Individual
4	IPDst-PortDst_IPDst-SrcBytes	0.7836	0.6221	True	True
10	IPDst-PortDst_IPDst-SrcBytes	0.9964	0.9938	True	True
11	IPDst-PortDst_PortSrc-PortDst	0.9970	0.9969	True	True
All	IPDst-PortDst_IPDst-SrcBytes	0.9901	0.9821	True	True

B. CTU-13

For the CTU-13 dataset, we summarized the results into Tables IV, V, and VI for the *MAX*, *MIN*, and *AVG* combination strategies. These Tables follow the same structure introduced in the previous subsection.

From the obtained results, we can see that, in all cases, the combination improved both over the SOTA as well as over the individual attribute correlations that compose it. It is worth pointing out the improvement in the performance of Scenario 4 of 0.15 to 0.20 over SOTA with all merging strategies while keeping the remaining scenarios above 0.99. We can also observe the presence of the *Destination IP Address* $\longrightarrow$ *Destination Port* attribute correlation in every H.P.C as well as the presence of the *Destination IP Address* $\longrightarrow$ *Bytes Sent* attribute correlation in all but one scenario across the experiments with all combination strategies.

C. UNSW-NB15

For the UNSW-NB15 dataset, we summarized the results into Tables VII, VIII, and IX for the *MAX*, *MIN*, and *AVG* combination strategies. These Tables follow the same structure introduced in the previous subsections.

From the obtained results, we can observe that there is only one case where the combination does not improve over its individual member – scenario 1, when using the *MIN* combination strategy – while it improves over both its individual members as well as SOTA in all the remaining cases. For this dataset, there is no single attribute correlation that is significantly more frequent than the rest, but we can highlight a consistent appearance of the *Destination IP Address* $\longrightarrow$ *Destination Port* attribute correlation as well as that of either the *Source IP Address* $\longrightarrow$ *Bytes Sent* or the *Destination IP Address* $\longrightarrow$ *Bytes Sent*.

TABLE V
SUMMARY OF RESULTS RUNNING MIMC OVER THE CTU-13 DATASET WITH THE MIN COMBINATION STRATEGY

Scenario	H.P.C	MIMC Score	SOTA Score	>SOTA	>Individual
4	IPDst-PortDst_IPSrc-PortDst_IPSrc-SrcBytes	0.8468	0.6221	True	True
10	IPDst-PortDst_IPDst-SrcBytes	0.9972	0.9938	True	True
11	IPDst-PortDst_IPDst-SrcBytes_IPSrc-SrcBytes	0.9978	0.9969	True	True
All	IPDst-PortDst_IPDst-SrcBytes_IPSrc-IPDst_ IPSrc-SrcBytes_PortSrc-PortDst	0.9919	0.9821	True	True

TABLE VI
SUMMARY OF RESULTS RUNNING MIMC OVER THE CTU-13 DATASET WITH THE AVG COMBINATION STRATEGY

Scenario	H.P.C	MIMC Score	SOTA Score	>SOTA	>Individual
4	IPDst-PortDst_IPDst-SrcBytes	0.7827	0.6221	True	True
10	IPDst-PortDst_IPDst-SrcBytes	0.9966	0.9938	True	True
11	IPDst-PortDst_IPDst-SrcBytes	0.9969	0.9969	True	True
All	IPDst-PortDst_IPDst-SrcBytes	0.9902	0.9821	True	True

TABLE VII
SUMMARY OF RESULTS RUNNING MIMC OVER THE UNSW-NB15 DATASET WITH THE MAX COMBINATION STRATEGY

Scenario	H.P.C	MIMC Score	SOTA Score	>SOTA	>Individual
1	IPDst-SrcBytes_IPSrc-IPDst_IPSrc-SrcBytes	0.8546	0.7577	True	True
2	IPDst-PortDst_IPSrc-PortDst	0.9737	0.9732	True	True
3	IPDst-PortDst_IPSrc-SrcBytes	0.7982	0.7513	True	True
4	IPDst-PortDst_IPDst-SrcBytes	0.8257	0.6945	True	True
All	IPDst-SrcBytes_IPSrc-SrcBytes	0.8796	0.8733	True	True

TABLE VIII
SUMMARY OF RESULTS RUNNING MIMC OVER THE UNSW-NB15 DATASET WITH THE MIN COMBINATION STRATEGY

Scenario	H.P.C	MIMC Score	SOTA Score	>SOTA	>Individual
1	IPDst-SrcBytes_IPSrc-IPDst_IPSrc-SrcBytes	0.7789	0.7577	True	False
2	IPDst-SrcBytes_IPSrc-IPDst	0.9721	0.9732	True	True
3	IPDst-PortDst_IPSrc-IPDst_IPSrc-SrcBytes	0.799	0.7513	True	True
4	IPDst-SrcBytes_IPSrc-SrcBytes	0.8376	0.6945	True	True
All	IPDst-SrcBytes_IPSrc-IPDst_IPSrc-SrcBytes_ PortSrc-PortDst	0.9033	0.8733	True	True

V. CONCLUSION & FUTURE WORK

The proposed method, MIMC, shows an increase in performance over SOTA techniques of micro-cluster-based anomaly detection. We notice that this improvement comes not only over entire datasets, as evaluated by similar SOTA techniques, but also over divisions of these datasets, which shows its ability to work with smaller datasets. By exploring different attribute correlations, MIMC is able to better detect anomalies in network data and produce more accurate results.

Given the large number of possible configurations in which MIMC can be run, we suggest a standard configuration that performs well in most cases. Given its consistent presence as part of the Highest Performing Combination of Attribute Correlation (H.P.C) in the experiments, we suggest combining data instances using the *Source Port* $\longrightarrow$ *Destination Port* and the *Destination IP Address* $\longrightarrow$ *Destination Port* attribute correlations. Given its consistency in outperforming SOTA and its individual members, we recommend using the *MAX* combination strategy, although any of the available strategies yields satisfactory results. It is worth mentioning that each application could have a set of attribute correlations and another combination strategy that could perform better depending on the data source properties.

The detection of anomalies is a primary method for identifying malicious behaviours in production systems. State-of-the-art techniques make use of micro-cluster-based detection over attributes extracted from application logs that record the traffic between hosts. In this study, we have shown that these SOTA techniques fail to account for some of the available information, leaving behind the possibility to detect a wider variety of attacks. We have also shown that by combining multiple instances of micro-cluster-based detection, we are able to improve on the results yielded by the SOTA in the majority of the scenarios even when a smaller number of log entries is available. This overcomes one of the limitations of the SOTA methods on sparse and small datasets. The proposed method, MIMC, stands on the two main aspects presented here: higher number of attribute correlations and combination strategies.

Future work will explore new attribute correlations, test new combination strategies, and evaluate the performance of MIMC over different datasets with additional attack scenarios.

ACKNOWLEDGEMENT

This research was enabled by the support of the NSERC Alliance Grant. The first author gratefully acknowledges the support by the province of Nova Scotia. The research is conducted as part of the Dalhousie NIMS Lab[1].

REFERENCES

[1] L. Akoglu, H. Tong, and D. Koutra. Graph based anomaly detection and description: a survey. *Data Mining and Knowledge Discovery*, 29:626—-688, 2015.

[2] E. Balkanli, N. Zincir-Heywood, and M. Heywood. Feature selection for robust backscatter ddos detection. In *2015 IEEE 40th Local Computer Networks Conference Workshops (LCN Workshops)*, pages 611–618, 2015.

[3] Siddharth Bhatia, Bryan Hooi, Minji Yoon, Kijung Shin, and Christos Faloutsos. Midas: Microcluster-based detector of anomalies in edge streams. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 3242–3249, 2020.

[4] Rafael Copstein, Bradley Nibblet, Andrew Johnston, Jeff Schwartzentruber, Malcolm Heywood, and Nur Zincir-Heywood. Towards anomaly detection via multiple instances of micro-cluster detection. In *Proceedings of the seventh IEEE CSNET Cyber Security in Networking Conference*, 2023.

[5] Sebastian Garcia, Martin Grill, Jan Stiborek, and Alejandro Zunino. An empirical comparison of botnet detection methods. *computers & security*, 45:100–123, 2014.

[6] Sebastian Garcia, Agustin Parmisano, and Maria Jose Erquiaga. IoT-23: A labeled dataset with malicious and benign IoT network traffic, January 2020. More details here https://www.stratosphereips.org /datasets-iot23.

[7] F. Haddadi, D. Runkel, N. Zincir-Heywood, and M. Heywood. On botnet behaviour analysis using gp and c4.5. GECCO Comp '14, page 1253–1260, New York, NY, USA, 2014. Association for Computing Machinery.

[8] Adarsh Kulkarni, Priya Mani, and Carlotta Domeniconi. Network-based anomaly detection for insider trading. *arXiv preprint arXiv:1702.05809*, 2017.

[9] Kabul Kurniawan, Andreas Ekelhart, Elmar Kiesling, Dietmar Winkler, Gerald Quirchmayr, and A Min Tjoa. Vlograph: a virtual knowledge graph framework for distributed security log analysis. *Machine Learning and Knowledge Extraction*, 4(2), 2022.

[10] Ling Lin and Jinshan Su. Anomaly detection method for sensor network data streams based on sliding window sampling and optimized clustering. *Safety science*, 118:70–75, 2019.

[1] https://projects.cs.dal.ca/projectx/

TABLE IX
SUMMARY OF RESULTS RUNNING MIMC OVER THE UNSW-NB15 DATASET WITH THE AVG COMBINATION STRATEGY

Scenario	H.P.C	MIMC Score	SOTA Score	>SOTA	>Individual
1	IPSrc-IPDst_IPSrc-SrcBytes	0.8517	0.7577	True	True
2	IPSrc-IPDst_PortSrc-PortDst	0.9732	0.9732	True	True
3	IPDst-PortDst_IPSrc-IPDst	0.8016	0.7513	True	True
4	IPDst-SrcBytes_PortSrc-PortDst	0.8271	0.6945	True	True
All	IPDst-SrcBytes_IPSrc-SrcBytes	0.8795	0.8733	True	True

[11] Misael Mongiovi, Petko Bogdanov, Razvan Ranca, Evangelos E Papalexakis, Christos Faloutsos, and Ambuj K Singh. Netspot: Spotting significant anomalous regions on dynamic networks. In *Proceedings of the 2013 Siam international conference on data mining*, pages 28–36. SIAM, 2013.

[12] Nour Moustafa and Jill Slay. Unsw-nb15: a comprehensive data set for network intrusion detection systems (unsw-nb15 network data set). In *2015 military communications and information systems conference (MilCIS)*, pages 1–6. IEEE, 2015.

[13] Caleb C Noble and Diane J Cook. Graph-based anomaly detection. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 631–636, 2003.

[14] Mandana Saebi, Jian Xu, Lance M Kaplan, Bruno Ribeiro, and Nitesh V Chawla. Efficient modeling of higher-order dependencies in networks: from algorithm to application for anomaly detection. *EPJ Data Science*, 9(1):15, 2020.

[15] Blake E Strom, Andy Applebaum, Doug P Miller, Kathryn C Nickels, Adam G Pennington, and Cody B Thomas. Mitre att&ck: Design and philosophy. In *Technical report*. The MITRE Corporation, 2018.

[16] Takeaki Uno, Hiroki Maegawa, Takanobu Nakahara, Yukinobu Hamuro, Ryo Yoshinaka, and Makoto Tatsuta. Micro-clustering: finding small clusters in large diversity. *arXiv preprint arXiv:1507.03067*, 2015.

[17] Xiaolan Wang, Md Manjur Ahmed, Mohd Nizam Husen, Hai Tao, and Qian Zhao. Dynamic micro-cluster-based streaming data clustering method for anomaly detection. In *Soft Computing in Data Science: 7th International Conference, SCDS 2023, Virtual Event, January 24–25, 2023, Proceedings*, pages 61–75. Springer, 2023.