Multivariate Metrics of Normal and Anomalous Network Behaviors

Nong Ye
School of Computing, Informatics, and Decision Systems Engineering
Arizona State University
Tempe, Arizona, USA
nongye@asu.edu

Douglas Montgomery
ITL/Advanced Network Technologies Division
National Institute of Standards and Technology
Gaithersburg, Maryland, USA
dougm@nist.gov

Kevin Mills
ITL/Advanced Network Technologies Division
National Institute of Standards and Technology
Gaithersburg, Maryland, USA
kevin.mills@nist.gov

Mark Carson
ITL/Advanced Network Technologies Division
National Institute of Standards and Technology
Gaithersburg, Maryland, USA
mark.carson@nist.gov

***Abstract*—Detecting network anomalies is a fundamental part of day to day operations for Internet Service Providers and enterprises to maintain the efficiency and reliability of computer networks. Network anomaly detection is based on data characteristics of normal and anomalous network behaviors. Although many existing studies report univariate data characteristics of normal and anomalous network behaviors, there are few studies on multivariate data characteristics of normal and anomalous network behaviors. The goal of this study is to investigate multivariate data characteristics of normal and anomalous network behaviors using the Partial-Value Association Discovery (PVAD) algorithm. This paper illustrates the use of the PVAD algorithm to analyze network flow data of a medium size enterprise under the normal condition and an anomalous condition and reveal multivariate data characteristics of the normal and anomalous network flows in the form of multivariate data associations.**

Keywords—network anomaly detection; multivariate data characteristics of network flows; data mining

I. INTRODUCTION

Network anomalies including cyber attacks and network failures (e.g., a network outage affecting Comcast, Spectrum, Verizon, Cox and RCN across the country on November 6, 2017 [1]) have occurred consistently for decades and become commonplace in computer networks today. The rapidly emerging IoT (Internet of Things) along with many security-weak devices in IoT expands the space that hackers can exploit and utilize. Detecting network anomalies in a timely fashion is a fundamental part of day to day operations for Internet Service Providers and enterprises to maintain the efficiency and reliability of computer networks.

Current industry practices of signature-based network intrusion detection (SNID) have continually proved insufficient to address the ever-evolving variety of new attacks. New "zero-day" attacks typically go many months before being discovered and often remain a viable means of exploit for years even after discovery [2]. Network anomaly detection aims at detecting anomalies which manifest large differences from profiles of normal network behaviors, including new "zero-day" attacks [3, 4]. Network anomalies may include both malicious network attacks and non-malicious but anomalous network behaviors (e.g., network errors/failures, and non-malicious but unusual user behaviors such as those in response to new software or software upgrades). Hence, we consider three categories of network behaviors: normal, anomalous (non-malicious user behaviors and network errors/failures), and malicious.

Current practices by experienced network operators for identifying and diagnosing network anomalies (e.g. monitoring plots of network traffic) are mostly ad hoc [5], require persistent attention, and do not scale up to massive amounts of network traffic. This drives the high demand for automated network anomaly detection [6, 7]. Network anomaly detectors (NADs) today are ad hoc in that they are mostly driven by the application of machine learning/data mining (ML/DM) techniques which NADs employ to learn profiles of normal network behaviors from network traffic data and detect deviations from normal profiles as network anomalies. Because current ML/DM techniques lack the capability of explicitly identifying multivariate data associations of network behaviors and partial-value variable relations (i.e., relations that exist for certain but not all values of variables), NADs based on current ML/DM techniques miss the bigger, complex picture of normal, anomalous, and malicious network behaviors and generate a large number of alerts including many false positives. Due to the lack of the accurate and robust performance by individual NADs, security professionals in every industry (i.e., service providers, public sector, retail, manufacturing, utilities, healthcare, transportation, and finance) report that they deploy many security products from many vendors to monitor their computer networks [8]. This fragmented and multiproduct security approach exponentially increases the number of alerts that resource-strapped security teams must review, and hinders an organization's ability to manage network anomalies including security threats [8].

The goal of our work is to investigate multivariate data characteristics of normal, anomalous and malicious network

behaviors using the Partial-V Association Discovery (PVAD) algorithm [9-11]. This paper illustrates the use of the PVAD algorithm to analyze network flow data of a medium size enterprise under the normal condition and an anomalous condition and reveal multivariate data characteristics of the normal and anomalous network behaviors in the form of multivariate data associations. Hence, this study demonstrates a method of identifying multivariate data characteristics of normal and anomalous network behaviors which can be included in NADs to enhance their accuracy and robustness.

II. Network Flow Data And Analysis

The following two data sets were collected at a medium size enterprise with approximately 5,000 users and 30,000 devices on computer networks.

1) Normal TCP flow data between the Internet outside the enterprise and the enterprise servers which are accessible to public on the Internet (i.e., between the Internet and the enterprise's public servers), collected on September 1, 2014 when there were no network anomalies or cyber attacks.

2) Anomalous TCP flow data between the Internet and the enterprise's public servers which were identified using SNORT (https://www.snort.org), collected on one day in January 2014. In this data set, the destination port covered each of ports 1-65535 about 6 times between the source and destination IP addresses within the enterprise. This resulted from the enterprise using port scanners to scan its own ports for security scans at the beginning of the year in January. There were no replies to those port scans because they were dropped by the firewall. These part scans are not normal network behaviors of the enterprise's users and are considered anomalies.

A network session between a source host and a destination host for a certain network application is defined by the source IP address, the destination IP address, the source port, the destination port, and the protocol (e.g., TCP or UDP), and contains a sequence of packets called a flow. Flow data for a flow captures features of packets in each flow [12, 13]. The two data sets contain bi-flow data which contain features of packets flowing in both directions from the source to the destination and from the destination to the source in each flow.

In Data Set 1 there are 78,131 flows and 78,131 data records for those flows, respectively. In Data Set 2, there are 412,980 data records for 412,980 flows, respectively.

Table I lists nine data fields of network flows that we used in this study. The data fields have both categorical variables (i.e., source and destination ports) and numeric variables (i.e., Duration, Source-to-Destination Packets, Source-to-Destination Bytes, Destination-to-Source Packets, Destination-to-Source Bytes, and Goelocation Frequency). It is challenging to determine relations of multiple variables involving both categorical variables and numeric variables. However, the PVAD algorithm can handle both categorical variables and numeric variables to determine associations of variable values. In Step 1 of the PVAD algorithm, we plotted the sorted values of each numeric variable, identified data clusters, and used data clusters to define categorical values of the numeric variable, and thus transformed the numeric variable into a categorical variable with categorical values, as shown in Figure 1 for Geolocation Frequency (x9).

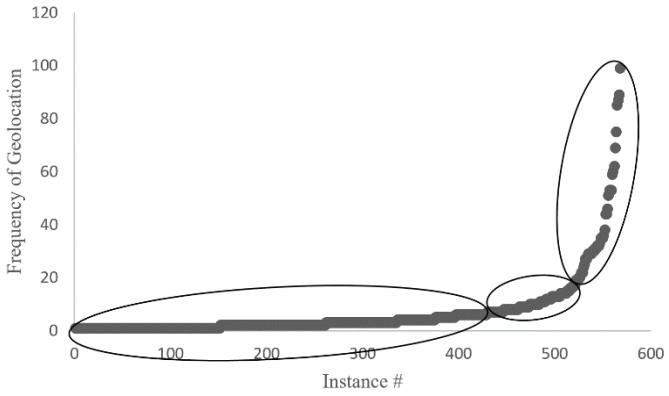

Fig. 1. The plot of the sorted values of Geolocation Frequency with the illustration of data clusters.

Although the two data fields, Source Port and Destination Port, are categorical variables, each has a large number of categorical values. To reduce the number of categorical values for Source Port, we kept system ports, but consolidated non-system ports. The port numbers in the range from 0 to 1023 are the system ports. The port number greater than 1023 are non-system ports. The field, Source Port, has only seven system ports in the data set: 20, 21, 22, 25, 53, 80, and 443, and we kept these system ports. For non-system ports of Source Port, we computed the frequency in which each non-system port appeared in the data set, and determined two categorical values: frequency $\leqslant$ 4, and frequency > 4. Similarly, we transformed the field, Destination Port, to have the categorical values as shown in Table I. A value gap between two categorical values, e.g., values greater than 2.017 and smaller than 2.02 in the gap of [0.076, 2.017] and [2.02, 31.716] for x2, means that values in the gap do not appear in the data set.

TABLE I. Categorical Variables of Data Sets 1 and 2

Data Field (Variable)	**Categorical Values**
Hour (x1)	24 hours
Duration (x2)	[0, 0.025], [0.026, 0.075], [0.076, 2.017], [2.02, 31.716], [31.74, 121.547], [125.272, 237.811], [271.055, 654.364], [707.653, 748.804], [751.86, 1655.325], [2159.123, 81327.16]
Source Port (x3)	20, 21, 22, 25, 53, 80, 443, NonSystem > 4, NonSystem ≤ 4
Destination Port (x4)	RegisteredSystem > 3, Registered > 3, Registered ≤ 3, Unregistered > 3, Unregistered ≤ 3
Source-to-Destination Packets (x5)	[1, 4], [5, 10], [11, 16], [17, 63], [64, 158], [161, 745], [884, 6017]
Source-to-Destination Bytes (x6)	[20, 52], [56, 143], [144, 240], [241, 593], [594, 799], [800, 1138], [1139, 1544], [1545, 2678], [2679, 2679], [2682, 5354], [5358, 8022], [8026, 29096], [29128, 69370], [69402,

	227566], [307752, 20482284]
Destination-to-Source Packets (x7)	[0, 2], [3, 8], [9, 14], [15, 48], [49, 105], [106, 173], [174, 204], [208, 256], [266, 307], [312, 1595]
Destination-to-Source Bytes (x8)	[0, 0], [20, 141], [143, 143], [144, 477], [478, 478], [479, 588], [589, 592], [593, 593], [594, 648], [649, 649], [650, 764], [765, 1009], [1010, 3005], [3026, 5357], [5364, 5381], [5391, 5429], [5430, 7980], [7986, 8094], [8106, 10682], [10688, 10826], [10842, 19429], [19461, 19626], [19640, 20129], [20137, 82848], [83229, 549241], [586818, 7267368]
Frequency of Source-Destination Geolocation (x9)	[1, 6], [7, 17], [19, 55033]

Steps 2 and 3 of the PVAD algorithm discover and consolidate associations of variable values, and each association is in the form of $X = A \rightarrow Y = B$, where X and Y are the vectors of one or more variables, A and B are the values of X and Y, respectively, $X = A$ are called conditional variables' values (CV), $Y = B$ are called associative variables' values (AV), and the co-occurrence ratio (cr) of each association is not smaller than α as follows:

$$cr(X = A \rightarrow Y = B) = \frac{N_{X=A \,\&\, Y=B}}{N_{X=A}} \geq \alpha \qquad (1)$$

where $N_{X=A,\ Y=B}$ is the number of supporting data records containing both $X = A$ and $Y = B$, and $N_{X=A}$ is the number of data records containing $X = A$. α is set to a value in the range of (0, 1] and is close or equal to 1. In this study, we set $\alpha = 0.95$.

In this study, we used the PVAD algorithm to identify 1-to-1 associations, 2-to-1 associations, …, and 8-to-1 associations. Table II lists the 8-to-1 associations with the number of supporting data records ⩾ 2000 for Data Set 1. Table III lists the 1-to-1 associations with the number of supporting instances ⩾ 20000 for Data Set 1.

TABLE II. 8-TO-1 ASSOCIATIONS WITH THE NUMBER OF SUPPORTING DATA RECORDS ⩾ 2000 FOR DATA SET 1

Association	**# of Supporting Data Records**
x1=8, x3=Nonsystem≤4, x4=RegisteredSystem>3, x5=[11, 16], x6=[1139, 1544], x7=[9, 14], x8=[19461, 19626], x9=[19, 55033]→x2=[0, 0.025]	2233
x1=8, x3=Nonsystem≤4, x4=RegisteredSystem>3, x5=[5, 10], x6=[241, 593], x7=[3, 8], x8=[650, 764], x9=[19, 55033]→x2=[0, 0.025]	6093
x2=[0, 0.025], x3=443, x4=RegisteredSystem>3, x5=[5, 10], x6=[2679, 2679], x7=[3, 8], x8=[1010, 3005], x9=[19, 55033]→x1=7	2462
x2=[0, 0.025], x3=Nonsystem≤4, x4=RegisteredSystem>3, x5=[11, 16], x6=[1139, 1544], x7=[9, 14], x8=[19461, 19626], x9=[19, 55033]→x1=8	2233
x2=[0, 0.025], x3=Nonsystem≤4, x4=RegisteredSystem>3, x5=[5, 10], x6=[241, 593], x7=[3, 8], x8=[650, 764], x9=[19, 55033]→x1=8	6093
x2=[0, 0.025], x3=Nonsystem≤4, x4=RegisteredSystem>3, x5=[5, 10], x6=[800, 1138], x7=[3, 8], x8=[5364, 5381], x9=[19, 55033]→x1=7	5518
x2=[0, 0.025], x3=Nonsystem≤4, x4=RegisteredSystem>3, x5=[5, 10], x6=[800, 1138], x7=[3, 8], x8=[5391, 5429], x9=[19, 55033]→x1=7	2472
x2=[0, 0.025], x3=Nonsystem≤4, x4=Unregistered≤3, x5=[5, 10], x6=[241, 593], x7=[3, 8], x8=[594, 648], x9=[19, 55033]→x1=7	7359

TABLE III. 1-TO-1 ASSOCIATIONS WITH THE NUMBER OF SUPPORTING DATA RECORDS ⩾ 20000 FOR DATA SET 1

Association	**# of Supporting Data Records**
x1=8→x4=RegisteredSystem>3	22119
x1=8→x3=Nonsystem≤4	22133
x1=8→x9=[19, 55033]	22262
x6=[241, 593]→x9=[19, 55033]	25186
x6=[241, 593]→x7=[3, 8]	25406
x6=[241, 593]→x5=[5, 10]	25525
x1=7→x9=[19, 55033]	35029
x2=[0, 0.025]→x9=[19, 55033]	40498
x5=[5, 10]→x7=[3, 8]	**52655**
x5=[5, 10]→x9=[19, 55033]	53936
x7=[3, 8]→x9=[19, 55033]	56563
x4=RegisteredSystem>3→x9=[19, 55033]	61616
x3=Nonsystem<=4→x9=[19, 55033]	62732

The 1-to-1 associations, …, 8-to-1 associations for Data Set 1 reveal multivariate data characteristics of normal network flows in the enterprise at that time. For example, the 1-to-1 association in Table III highlighted in bold indicates the association of source packets and destination packets in that source packets in the range [5, 10] are associated with destination packets in the range of [3, 8]. This 1-to-1 association has 52655 supporting data records or 67.39% (52655/78131) of data records in Data Set 1. Note from Table III that 71% of data records in Data Set 1 have x5 = [5, 10], and that 74% of data records in Data Set 1 have x7 = [3, 8]. Thus, 52.54% (71% * 74%) of data records in the data set are expected to have x5 = [5, 10] and x7 = [3, 8] by probability. 67.39% of data records in the data set actually have x5 = [5, 10] and x7 = [3, 8] indicates the association of x5 = [5, 10] with x7 = [3, 8] exists not by probability but due to a particular nature of flows as information sent out in destination packets is the response to requests in source packets. The small number of source packets, the small number of destination packets and their association may be attributed to many flows involving brief exchanges between a source and a destination during the establishment of a network session but without the need of transmitting large amounts of data in the established network session.

For another example, the last 8-to-1 association in Table II has the largest number of supporting data records among all the 8-to-1 associations, and this association reveals that the flows with the short duration of [0, 0.025], the less frequent non-system source port (frequency ⩽ 4), the less frequent

unregistered non-system destination port (frequency ⩽ 3), the source packets in the range of [5, 10], the destination packets in the range of [241, 593], the source bytes in the range of [3, 8], the destination bytes in the range of [594, 648], and a frequent pair of the IP source and destination addresses mostly occurred at 7 AM.

The PVAD algorithm was also used to analyze data in Data Set 2. The associations with the largest numbers of supporting data records reveal the associations of the following variable values:

x2=[0, 0.025],

x3=Nonsystem>4,

x4=RegisteredOthers&UnregisteredOthers>3,

x5=[1, 4],

x6=[20, 52],

x7=[0, 2],

x8=[0, 0],

x9=[19, 55033],

which occur mostly at 11 am (x1=11).

For Data Set 2, multivariate characteristics in the form of associations reveal the data characteristics of flows for port scans with a short duration, small source packets, no response packets over a wide range of destination ports.

Multivariate association of port scans are different from the dominant multivariate associations of the following attribute values in normal network flows in Data Set 1 such as the following:

- x5 = [5, 10], x7 = [3, 8];
- x1=7, x2=[0, 0.025], x3=Nonsystem⩽4, x4=Unregistered ⩽3, x5=[5, 10], x6=[241, 593], x7=[3, 8], x8=[594, 648], x9=[19, 55033];
- x1=7, x2=[0, 0.025], x3=Nonsystem ⩽ 4, x4=RegisteredSystem>3, x5=[5, 10], x6=[800, 1138], x7=[3, 8], x9=[19, 55033].

The above set of data associations as the dominant multivariate data characteristics of normal network flows and the dominant data associations of anomalous network flows (e.g., the association of x2=[0, 0.025], x3=Nonsystem>4, x4=RegisteredOthers&UnregisteredOthers>3, x5=[1, 4], x6=[20, 52], x7=[0, 2], x8=[0, 0], x9=[19, 55033] of the port scan flows) can be used to detect network anomalies by monitoring network flow data and keep track of percentages of those associations appear in network flow data. When data associations of normal network flows stay dominant, no network anomalies are detected. When data associations of normal network flows become less dominant and/or data associations of anomalous network flows become more dominant, network anomalies may be detected.

III. SUMMARY

Using the data sets of normal and anomalous TCP flow data collected at the enterprise, the study described in this paper illustrates how the PVAD algorithm is used to analyze these data sets and establish multivariate data characteristics of normal and anomalous network flow data. For example, this study indicates the important role that source and destination packets and bytes can play an important role in network anomaly detection in addition to source and destination ports which are often used for network anomaly detection. The methodology of using the PVAD algorithm to analyze network flow data and establish multivariate data associations of different network behaviors can be used by any organization to analyze its network data and establish data characteristics as metrics used to build NADs for the organization.

REFERENCES

[1] Newman LH (2017) How a Tiny Error Shut Off the Internet for Parts of the US. https://www.wired.com/story/how-a-tiny-error-shut-off-the-internet-for-parts-of-the-us/.

[2] RAND Corporation (2017) Zero Days, Thousands of Nights: The Life and Times of Zero-Day Vulnerabilities and Their Exploits.

[3] Ye N (2008) Secure Computer and Network Systems: Modeling, Analysis and Design (John Wiley & Sons, London, UK).

[4] Chandola V, Banerjee A, Kumar V (2009) Anomaly detection: A survey. ACM Computing Surveys 41(3): Article 15.

[5] Barford P, Kline J, Plonka D, Ron A (2002) A signal analysis of network traffic anomalies. In Proceedings of IMW'02, pp. 71-82.

[6] Inampudi K (2017) How data science can boost network operations. https://www.lightreading.com/automation/how-data-science-can-boost-network-operations/a/d-id/736773.

[7] Abhishek C (2017). Anomaly detection market is expected to grow worth 4.45 Billion USD by 2022). https://www.linkedin.com/pulse/anomaly-detection-market-expected-grow-worth-445-billion-c-abhishek.

[8] Cisco 2017 Midyear Security Report (2017) http://go.flashpoint-intel.com/docs/Flashpoint-Cisco-2017-Midyear-Cybersecurity-Report.

[9] Ye, N, Fok TY (2019) Learning partial-value variable associations. In Proceedings of 2019 4th International Conference on Big Data and Computing, Guangzhou, China, May 10-12, 2019.

[10] Ye N (2017) Analytical techniques for anomaly detection through features, signal-noise separation and partial-value associations. In Proceedings of Machine Learning Research, Vol. 77, pp. 1-10.

[11] Ye N (2017) The partial-value association discovery algorithm to learn multi-layer structural system models from system data. IEEE Transactions on Systems, Man, and Cybernetics: Systems 47(12): pp. 3377-3385.

[12] Dainotti A, Pescape A, Ventre G (2006) A packet-level characterization of network traffic. In Proceedings of the 11th International Workshop on Computer-Aided Modeling, Analysis and Design of Communication Links and Networks, pp. 38-45, 2006.

[13] Sperotto A, Schaffrath G, Sadre R, Morariu C, Pras A, Stiller B (2010). An overview of IP flow-based intrusion detection. IEEE Communications Survey & Tutorials 12(3): 343-356.