Anomaly analysis on an open DNS dataset1

Aziz, B, Menychtas , N & Al Bazi, A

DOI 10.7287/peerj.preprints.27116v1

Publisher: Peerj preprints

Anomaly analysis on an open DNS dataset

Benjamin Aziz[1], **Nikolaos Menychtas**[2], **and Ammar Al-Bazi**[3]

[1]**School of Computing, University of Portsmouth, UK**
[2]**Zananet, Southampton, UK**
[3]**Coventry University, Coventry, UK**

Corresponding author:
Benjamin Aziz[1]

Email address: benjamin.aziz@port.ac.uk

ABSTRACT

The increasing availability of open data and the demand to understand better the nature of anomalies and the causes underlying them in modern systems is encouraging researchers to analyse open datasets in various ways. These include both quantitative and qualitative methods. We show here how quantitative methods, such as timeline, local averages and exponentially weighted moving average analyses, led in this work to the discovery of three anomalies in a large open DNS dataset published by the Los Alamos National Laboratory.

INTRODUCTION

Large datasets are becoming ever more available in open formats for various domains of technology driven by the aim of creating shared knowledge beyond the capabilities that a single organisation can generate. Such knowledge is valuable as it maintains and facilitates the operation of a robust, efficient and reliable IT infrastructure. As a result, the analysis and mining of large and open datasets has become, in recent times, an important and integral part of the research activities in successful IT teams, particularly within the scope of Cyber security research. In recent years, we have witnessed the arrival of large open Cyber security datasets, e.g. VCDB [23], CERT's Vulnerability Notes Database at Carnegie Mellon University [4], SecRepo [8], CAIDA [3] and LANL [7]), backed and maintained by reputable organisations.

In this short paper, we summarise the results of one such analytical exercise we performed on a large and open dataset containing Internet events, namely the Domain Name Service (DNS) dataset [5, 1] provided and maintained by the Los Alamos National Laboratory [6]. Our analysis follows three methods: a timeline analysis to understand whether there exist any gaps in the timeline, a local averages analysis, which identifies the server's average load in each timeline period, and the Exponentially Weighted Moving Average (EWMA) [16] analysis, which results in a control chart that monitors the progress of the DNS workload.

RELATED WORK

Anomaly analysis of computing and communication-related datasets using statistical methods such as the EWMA method is not a new idea and it has been researched and applied in literature on several occasions [24, 2, 12]. Viinikka and Debar [24], for example, presented an alert processing method based on EWMA control charts to summarise the behaviour of alert flows to meet a set of five objectives. These objectives included anomaly highlighting, decreasing operator load, reduction measurement and determination of suitable flows for monitoring and trend visualisation. Carter and Streilein [2], on the other hand, employed a probabilistic weighting method to the standard EWMA method to dynamically adjust parameterisation based on the probability of a given observation. Osanaiye, Alfa and Hancke [12] used the EWMA method to detect anomalous changes in the intensity of a jamming attack event. This is achieved by monitoring the packet inter-arrival feature of the received packets from sensor nodes. In [11] in 2002, Ye, Borror and Zhang used the EWMA method in three instances; for auto-correlated data, for uncorrelated data and for the standard deviation, to detect Denial-of-Service (DoS) attacks in computer networks, therefore becoming one of the earliest works that suggested the application of the EWMA method to computer intrusion detection.

Other statistical methods have also been applied to the analysis of computer networks traffic where for example in [15, 22], Polunchenko, Tartakovsky, Mukhopadhyay and Sokolov used four statistical methods: the CUmulative SUM (CUSUM) [13], the Shiryaev-Roberts (SR) [19, 17], the Shiryaev-Roberts-Pollak (SRP) [14] and the Shiryaev-Roberts-r (SR-r) [10] methods to rapidly detect anomalies in such traffic, where an anomaly is considered to be a change in the traffic. More recently, Sklavounos, Edoh and Plytas [20] used the EMWA and the CUSUM methods to detect instances of the Root-to-Local (R2L) attacks, where the attacker sends packets to some remote computer with the aim of exploiting its vulnerabilities and acquiring privileges as a local user. The proposed method is used in detecting shifts of the normal process of the TCP source bytes during operation, which could imply an R2L attack.

Finally, in [21], Soldo, Le and Markopoulou used the EWMA method as a spatio-temporal pattern prediction tool to predict future attack sources from past attack logs that contain attacker-victim history and interactions. This is then implemented as a blacklisting recommendation system.

THE LANL DNS DATASET

Our analysis focuses on the DNS datasaet [5], part of the ”Comprehensive, Multi-Source Cyber-Security Event” datasets published by the Los Alamos National Laboratory (LANL). The dataset represents 58 consecutive days of de-identified DNS lookup events collected from within LANL's corporate internal computer network. Each event, expressed as a row, has the metadata (time, source computer, computer resolved). Therefore, the events have a minimalistic set of metadata or information associated with them: the time at which the event occurred, a pseudo-identity of the computer issuing the query and a pseudo-identity of the computer the query was resolved to. The time of the events themselves starts at an unknown epoch of ”1” and uses a time resolution of ”1” second. An example representing three entries from this dataset is shown below [5]:

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27116v1 | CC BY 4.0 Open Access | rec: 14 Aug 2018, publ: 14 Aug 2018

```
31,C161,C2109
35,C5642,C528
38,C3380,C22841
```

The dataset is 812MB in size and spans over 40,821,591 records. The dataset can therefore be described as Big, and it was published back in 2015.

THE ANALYSIS APPROACH

Our approach in analysing the LANL DNS Dataset [5] was driven by the nature of the data included. This mainly suggested two streams of analysis: First analysis of the timeline and second analysis of the DNS server workload. More specifically, we carried out the following three analyses.

First Method: Timeline Analysis

The first method we used is the timeline analysis, to discover if there were any time gaps in the DNS server's readings that would divide the timeline of the readings into periods. We define a gap, as a period of inactivity that exceeds 24 hours. Other definitions are possible where the length of this period of inactivity would vary. Assuming there are g number of such gaps, we can divide a timeline $\mathscr{T}$ into n number of activity periods, where $n = g + 1$.

Second Method: Local Averages Analysis

The second analysis method we applied is a *local averages* analysis. More precisely, given a timeline $\mathscr{T}$ extending over the period from 0 to time t, and divided into n number of periods (in our case $n = 2$, where $g = 1$), then a local averages analysis will produce the set $\mathscr{A} = \{av_1, \ldots, av_n\}$ representing the averages for each of the periods over which $\mathscr{T}$ is divided. Each av_i value is calculated as the average of the number of DNS requests made over the i^{th} period.

Third Method: Exponentially Weighted Moving Average Analysis

We adopted the Exponentially Weighted Moving Average (EWMA) statistic [16] as the third analysis technique for the LANL DNS dataset. EWMAs are a kind of statistical control charts, a concept first proposed by Shewhart in 1931 [18]. Shewhart control charts have been widely used for decades. However, since these charts use only the information contained in the current sample observation, they are not efficient in detecting small process parameter changes. On the other hand, EWMAs are better in detecting small shifts [9] and average data in a way that gives less and less weight to the data as they are further removed in time.

The EWMA analysis produces two control limits that define the band of values for the Y-axis that are considered to be *normal* and therefore under control. These limits are the *Upper Control Limit* (UCL) and the *Lower Control Limit* (LCL), and are calculated based on the standard deviation σ value for the Y-axis. The main rationale in choosing this method as the third kind of analysis is to determine what is normal and what is abnormal processing load for the DNS server. This is determined by adjusting

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27116v1 | CC BY 4.0 Open Access | rec: 14 Aug 2018, publ: 14 Aug 2018

the distance at which the UCL and LCL limits are set, which in reality will be based on the history of data and past experience with the server's behaviour. In our case, we chose (as an example) to set the limits to be at $25 \times \sigma$.

OUR FINDINGS

The general timeline analysis is shown in Figure 1.

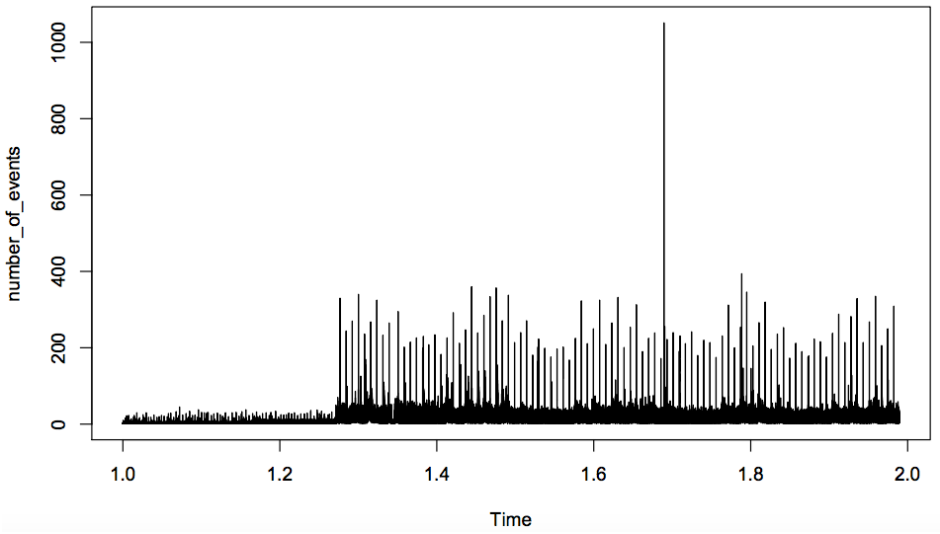

Figure 1. Timeline analysis of the LANL DNS dataset over the whole 58 days but not showing the first anomaly.

Below, we outline the findings we concluded from this analysis.

First Anomaly

The first anomaly we detected was the result of the application of the timeline analysis where we discovered the presence of a time gap of 77.1225 hours (i.e. 3 days, 5 hours, 7 minutes and 21 seconds) during which the DNS server readings were absent. This gap starts at time 2010062 (i.e. after approximately 23 days and 6 hours) and ends at time 2287703, inclusive. In the actual dataset, this gap is seen in-between these two rows:

```
2010061,C5948,C457
2287704,C12019,C1707
```

This indicates that the DNS server (or its configuration server) was taken down for this period, perhaps due to the presence of the second anomaly we discuss below. As a result, our timeline analysis divides the DNS dataset timeline $\mathscr{T}$ into two periods ($n = 2$) and one gap ($g = 1$).

Second Anomaly

The second anomaly we found was a result of the application of the local averages analysis, and it is related to the query processing ability of the DNS server over the whole period of the dataset. This analysis showed that the server in the first activity

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27116v1 | CC BY 4.0 Open Access | rec: 14 Aug 2018, publ: 14 Aug 2018

period of 23 days and 6 hours performed at a low workload, where the number of Queries it Processed per Second (or what is known as the *QPS* metric) was on average approximately 1.6. On the other hand, after time 2287704, when the server recovers from its downtime (first anomaly above), its QPS average rises in the second activity period to 14.8 over the last 31 days recorded in the dataset. We consider that the low QPS in the first period may have been caused by an earlier fault, misconfiguration or even an attack that prevented the server from processing queries at a normal workload.

Third Anomaly

We applied the EWMA statistic to the second activity period in the dataset's timeline, which was the last 31 days (or 2678400 seconds), as we consider this to be more of a normal workload period for the server. The resulting chart for this second period is shown in Figure 2.

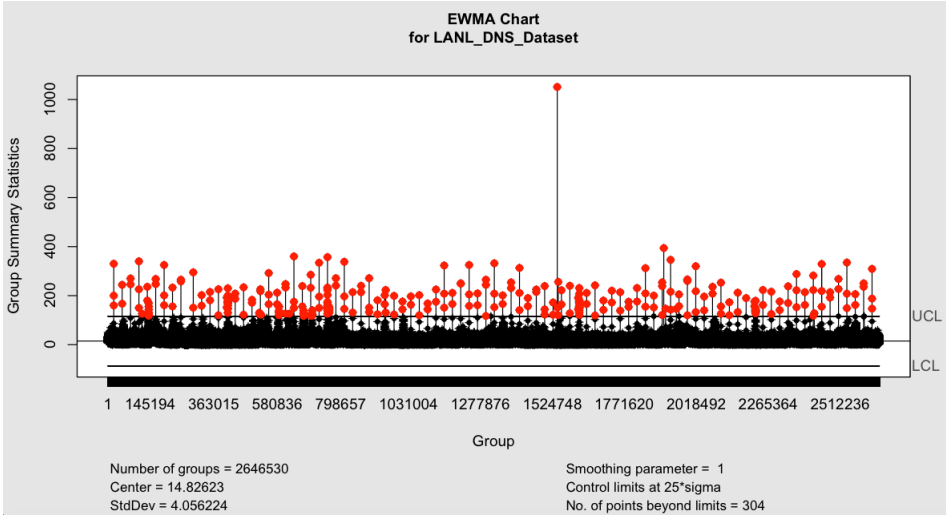

Figure 2. The EWMA chart for the last 31 days of the LANL DNS dataset for control limits of $25 \times \sigma$.

The black dots represent numbers of DNS requests per second that fall within the control limits, whereas the red dots represent cases where such numbers are outside of the UCL limit. The LCL limit here is a negative number, therefore it cannot be violated. As we mentioned earlier, one of the main benefits of an EWMA analysis is to determine whether a process is under control and highlight points that are outside of the normal control limits, therefore, prompt the administrators to further investigate those abnormal points.

Based on this approach, and by setting the limit to be at $25 \times \sigma$, we were able to discover points in time when the DNS server was not operating within the normal load. The classification is based on the choice of this limit. In our case, it confirmed that the "spike" in the number of queries processed by the DNS server at time 3906002 (i.e. on day 45, around the 5th hour) where 1051 queries were processed in that second, was indeed an unusual point in the chart. This spike is more than 70 times higher than the average QPS during this period and it is substantially higher than the next three highest spikes of 394, 360 and 357 queries per second occurring at times 4271510, 2998863 and

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27116v1 | CC BY 4.0 Open Access | rec: 14 Aug 2018, publ: 14 Aug 2018

3114002, respectively. Therefore, it does indicate some form of DoS attack or possible stress testing on the sever.

A different (but rather unusual) interpretation of the data would have been to choose the control limits sufficiently wide enough such that there would be no abnormal points, including the large spike at time 3906002. The choice of control limits is entirely dependant on the control procedures adopted by the organisation.

CONCLUSION AND FUTURE WORK

To conclude this short paper, we applied three analysis techniques to the LANL DNS open dataset in order to understand what kind of timeline and workload properties this dataset demonstrated. We were able to detect, as a result, three kinds of anomalies. The first indicated a period of time when the DNS server was not fully functional. The second anomaly showed that the server became non-functional (offline) for a short period of time, and finally, the third anomaly demonstrated an unusual spike in the number of queries that the server process in one second after it was restored.

In the future, we plan to apply other statistical analysis methods to the current dataset and to other datasets. We are also planning to investigate how to set the EWMA control limits in an automatic manner based on data mining techniques that utilise past experience to determine what normal load the server should be running at.

REFERENCES

[1] A. D. Kent (2015). Comprehensive, Multi-Source Cyber-Security Events.

[2] Carter, K. M. and Streilein, W. W. (2012). Probabilistic reasoning for streaming anomaly detection. In *2012 IEEE Statistical Signal Processing Workshop (SSP)*, pages 377–380.

[3] Center for Applied Internet Data Analysis (2018). CAIDA Data. Last accessed: 01.06.2018.

[4] CERT Coordination Center (2018). CERT Vulnerability Notes Database. Last accessed: 01.06.2018.

[5] Kent, A. D. (2015). Cybersecurity Data Sources for Dynamic Network Research. In *Dynamic Networks in Cybersecurity*. Imperial College Press.

[6] LANL (2018). Los Alamos National Laboratory: Cyber Security Science. `https://csr.lanl.gov/data/`. Accessed: 24-04-2018.

[7] Los Alamos National Laboratory (2018). Cyber Security Science Open Data Sets. Last accessed: 01.06.2018.

[8] Mike Sconzo (2018). SecRepo.com - Samples of Security Related Data. Last accessed: 01.06.2018.

[9] Montgomery, D. (2005). *Introduction to Statistical Quality Control*. John Wiley & Sons, Inc.

[10] Moustakides, G. V., Polunchenko, A. S., and Tartakovsky, A. G. (2011). A numerical approach to performance analysis of quickest change-point detection procedures. *Statistica Sinica*, 21(2):571–596.

[11] Nong, Y., Connie, B., and Yebin, Z. (2002). Ewma techniques for computer intrusion detection through anomalous changes in event intensity. *Quality and Reliability Engineering International*, 18(6):443–451.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27116v1 | CC BY 4.0 Open Access | rec: 14 Aug 2018, publ: 14 Aug 2018

[12] Osanaiye, O., Alfa, A., and Hancke, G. (2018). A statistical approach to detect jamming attacks in wireless sensor networks. *Sensors*, 18(6):1691.

[13] Page, E. S. (1954). Continuous inspection schemes. *Biometrika*, 41(1-2):100–115.

[14] Pollak, M. (1985). Optimal detection of a change in distribution. *The Annals of Statistics*, 13(1):206–227.

[15] Polunchenko, A. S., Tartakovsky, A. G., and Mukhopadhyay, N. (2012). Nearly Optimal Change-Point Detection with an Application to Cybersecurity. *Sequential Analysis*, 31(3):409–435.

[16] Roberts, S. W. (1959). Control Chart Tests Based on Geometric Moving Averages. *Technometrics*, 1(3):239–250.

[17] Roberts, S. W. (1966). A comparison of some control chart procedures. *Technometrics*, 8(3):411–430.

[18] Shewhart, W. (1931). *Economic Control of Quality of Manufactured Product*. Lancaster Press, Inc.

[19] Shiryaev, A. N. (1961). The problem of the most rapid detection of a disturbance in a stationary process. *Soviet Mathematics-Doklady*, 2:795–799.

[20] Sklavounos, D., Edoh, A., and Plytas, M. (2017). A Statistical Approach Based on EWMA and CUSUM Control Charts for R2L Intrusion Detection. In *2017 Cybersecurity and Cyberforensics Conference (CCC)*, pages 25–30.

[21] Soldo, F., Le, A., and Markopoulou, A. (2011). Blacklisting recommendation system: Using spatio-temporal patterns to predict future attacks. *IEEE Journal on Selected Areas in Communications*, 29(7):1423–1437.

[22] Tartakovsky, A. G., Polunchenko, A. S., and Sokolov, G. (2013). Efficient Computer Network Anomaly Detection by Changepoint Detection Methods. *IEEE Journal of Selected Topics in Signal Processing*, 7(1):4–11.

[23] VERIZON (2018). VERIS Community Database. Last accessed: 01.06.2018.

[24] Viinikka, J. and Debar, H. (2004). Monitoring ids background noise using ewma control charts and alert information. In Jonsson, E., Valdes, A., and Almgren, M., editors, *Recent Advances in Intrusion Detection*, pages 166–187, Berlin, Heidelberg. Springer Berlin Heidelberg.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27116v1 | CC BY 4.0 Open Access | rec: 14 Aug 2018, publ: 14 Aug 2018