

# 1 Anomaly analysis on an open DNS dataset

2 Benjamin Aziz<sup>1</sup>, Nikolaos Menychtas<sup>2</sup>, and Ammar Al-Bazi<sup>3</sup>

3 <sup>1</sup>School of Computing, University of Portsmouth, UK

4 <sup>2</sup>Zananet, Southampton, UK

5 <sup>3</sup>Coventry University, Coventry, UK

6 Corresponding author:

7 Benjamin Aziz<sup>1</sup>

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

## 9 ABSTRACT

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

## 17 INTRODUCTION

18 Large datasets are becoming ever more available in open formats for various domains  
19 of technology driven by the aim of creating shared knowledge beyond the capabilities  
20 that a single organisation can generate. Such knowledge is valuable as it maintains and  
21 facilitates the operation of a robust, efficient and reliable IT infrastructure. As a result,  
22 the analysis and mining of large and open datasets has become, in recent times, an  
23 important and integral part of the research activities in successful IT teams, particularly  
24 within the scope of Cyber security research. In recent years, we have witnessed the  
25 arrival of large open Cyber security datasets, e.g. VCDB [23], CERT's Vulnerability  
26 Notes Database at Carnegie Mellon University [4], SecRepo [8], CAIDA [3] and LANL  
27 [7]), backed and maintained by reputable organisations.

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

## 36 RELATED WORK

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

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

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

## 69 THE LANL DNS DATASET

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

80  
81 31, C161, C2109  
82 35, C5642, C528  
83 38, C3380, C22841  
84

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

## 87 THE ANALYSIS APPROACH

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

### 92 **First Method: Timeline Analysis**

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

### 99 **Second Method: Local Averages Analysis**

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

### 106 **Third Method: Exponentially Weighted Moving Average Analysis**

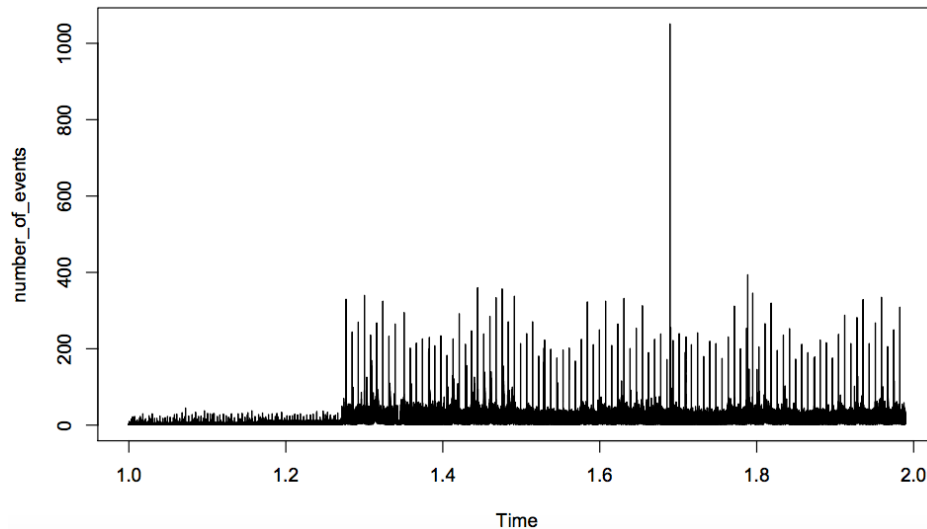
107 We adopted the Exponentially Weighted Moving Average (EWMA) statistic [16] as the  
108 third analysis technique for the LANL DNS dataset. EWMA's are a kind of statistical  
109 control charts, a concept first proposed by Shewhart in 1931 [18]. Shewhart control  
110 charts have been widely used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

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

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

## 124 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.

125

126

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

### 127 First Anomaly

128 The first anomaly we detected was the result of the application of the timeline analysis  
 129 where we discovered the presence of a time gap of 77.1225 hours (i.e. 3 days, 5 hours,  
 130 7 minutes and 21 seconds) during which the DNS server readings were absent. This  
 131 gap starts at time 2010062 (i.e. after approximately 23 days and 6 hours) and ends at  
 132 time 2287703, inclusive. In the actual dataset, this gap is seen in-between these two rows:

133

134 2010061, C5948, C457

135 2287704, C12019, C1707

136

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

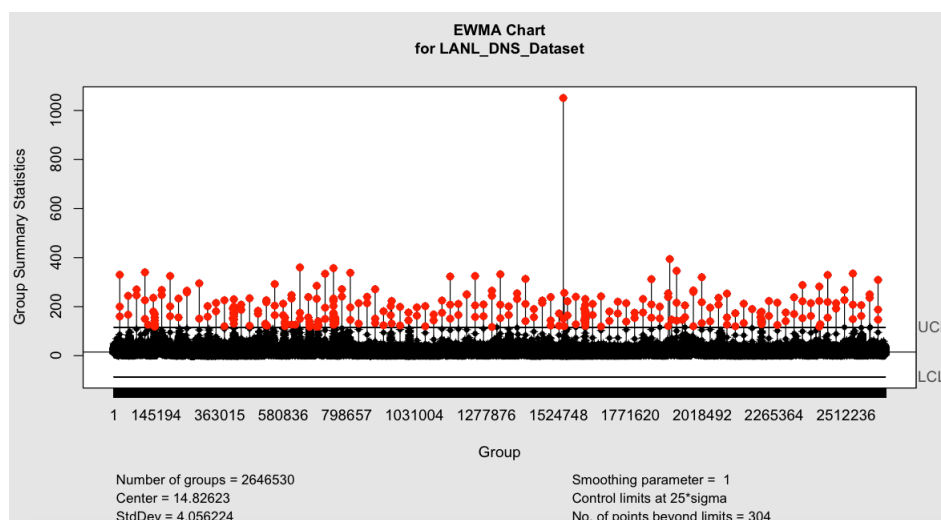
### 141 Second Anomaly

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

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

### 152 Third Anomaly

153 We applied the EWMA statistic to the second activity period in the dataset's timeline,  
 154 which was the last 31 days (or 2678400 seconds), as we consider this to be more of  
 155 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  $25 \times \sigma$ .

156  
 157 The black dots represent numbers of DNS requests per second that fall within the  
 158 control limits, whereas the red dots represent cases where such numbers are outside of  
 159 the UCL limit. The LCL limit here is a negative number, therefore it cannot be violated.  
 160 As we mentioned earlier, one of the main benefits of an EWMA analysis is to determine  
 161 whether a process is under control and highlight points that are outside of the normal  
 162 control limits, therefore, prompt the administrators to further investigate those abnormal  
 163 points.

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

172 3114002, respectively. Therefore, it does indicate some form of DoS attack or possible  
173 stress testing on the sever.

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

## 178 CONCLUSION AND FUTURE WORK

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

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

## 190 REFERENCES

- 191 [1] A. D. Kent (2015). Comprehensive, Multi-Source Cyber-Security Events.
- 192 [2] Carter, K. M. and Streilein, W. W. (2012). Probabilistic reasoning for streaming  
193 anomaly detection. In *2012 IEEE Statistical Signal Processing Workshop (SSP)*,  
194 pages 377–380.
- 195 [3] Center for Applied Internet Data Analysis (2018). CAIDA Data. Last accessed:  
196 01.06.2018.
- 197 [4] CERT Coordination Center (2018). CERT Vulnerability Notes Database. Last  
198 accessed: 01.06.2018.
- 199 [5] Kent, A. D. (2015). Cybersecurity Data Sources for Dynamic Network Research. In  
200 *Dynamic Networks in Cybersecurity*. Imperial College Press.
- 201 [6] LANL (2018). Los Alamos National Laboratory: Cyber Security Science. <https://csr.lanl.gov/data/>. Accessed: 24-04-2018.
- 202 [7] Los Alamos National Laboratory (2018). Cyber Security Science Open Data Sets.  
203 Last accessed: 01.06.2018.
- 204 [8] Mike Sconzo (2018). SecRepo.com - Samples of Security Related Data. Last  
205 accessed: 01.06.2018.
- 206 [9] Montgomery, D. (2005). *Introduction to Statistical Quality Control*. John Wiley &  
207 Sons, Inc.
- 208 [10] Moustakides, G. V., Polunchenko, A. S., and Tartakovsky, A. G. (2011). A numerical  
209 approach to performance analysis of quickest change-point detection procedures.  
210 *Statistica Sinica*, 21(2):571–596.
- 211 [11] Nong, Y., Connie, B., and Yebin, Z. (2002). Ewma techniques for computer intrusion  
212 detection through anomalous changes in event intensity. *Quality and Reliability  
213 Engineering International*, 18(6):443–451.

- 215 [12] Osanaiye, O., Alfa, A., and Hancke, G. (2018). A statistical approach to detect  
216 jamming attacks in wireless sensor networks. *Sensors*, 18(6):1691.
- 217 [13] Page, E. S. (1954). Continuous inspection schemes. *Biometrika*, 41(1-2):100–115.
- 218 [14] Pollak, M. (1985). Optimal detection of a change in distribution. *The Annals of*  
219 *Statistics*, 13(1):206–227.
- 220 [15] Polunchenko, A. S., Tartakovsky, A. G., and Mukhopadhyay, N. (2012). Nearly  
221 Optimal Change-Point Detection with an Application to Cybersecurity. *Sequential*  
222 *Analysis*, 31(3):409–435.
- 223 [16] Roberts, S. W. (1959). Control Chart Tests Based on Geometric Moving Averages.  
224 *Technometrics*, 1(3):239–250.
- 225 [17] Roberts, S. W. (1966). A comparison of some control chart procedures. *Technomet-*  
226 *rics*, 8(3):411–430.
- 227 [18] Shewhart, W. (1931). *Economic Control of Quality of Manufactured Product*.  
228 Lancaster Press, Inc.
- 229 [19] Shiryaev, A. N. (1961). The problem of the most rapid detection of a disturbance in  
230 a stationary process. *Soviet Mathematics-Doklady*, 2:795–799.
- 231 [20] Sklavounos, D., Edoh, A., and Plytas, M. (2017). A Statistical Approach Based  
232 on EWMA and CUSUM Control Charts for R2L Intrusion Detection. In *2017*  
233 *Cybersecurity and Cyberforensics Conference (CCC)*, pages 25–30.
- 234 [21] Soldo, F., Le, A., and Markopoulou, A. (2011). Blacklisting recommendation  
235 system: Using spatio-temporal patterns to predict future attacks. *IEEE Journal on*  
236 *Selected Areas in Communications*, 29(7):1423–1437.
- 237 [22] Tartakovsky, A. G., Polunchenko, A. S., and Sokolov, G. (2013). Efficient Computer  
238 Network Anomaly Detection by Changepoint Detection Methods. *IEEE Journal of*  
239 *Selected Topics in Signal Processing*, 7(1):4–11.
- 240 [23] VERIZON (2018). VERIS Community Database. Last accessed: 01.06.2018.
- 241 [24] Viinikka, J. and Debar, H. (2004). Monitoring ids background noise using ewma  
242 control charts and alert information. In Jonsson, E., Valdes, A., and Almgren, M.,  
243 editors, *Recent Advances in Intrusion Detection*, pages 166–187, Berlin, Heidelberg.  
244 Springer Berlin Heidelberg.