Analysis of malicious network flows using statistical natural laws

Pedro Fernandes[12]
Pedro.Fernandes@tus.ie
Séamus Ó Ciardhuáin[1]
Seamus.OCiardhuain@tus.ie
Mário Antunes[234]
mario.antunes@ipleiria.pt

[1] Department of Information Technology, Technological University of the Shannon, Moylish Park, V94 EC5T Limerick, Ireland
[2] Polytechnic of Leiria, School of Technology and Management Leiria, Portugal
[3] Computer Science and Communication Research Centre, Polytechnic of Leiria - Portugal
[4] INESC-TEC, CRACS, Porto - Portugal

Abstract

The detection of network flows carrying malicious data has been left to a set of techniques and tools based on machine learning (ML), with the disadvantage of requiring vast computational resources and sometimes low generalisation capacity when faced with new types of attack, known as a zero-day attack.

This paper describes the application of a model based on three statistical natural laws, Benford's, Stigler's and Zipf's Laws, as a model for detecting malicious flows extracted from network traffic, as well as the results obtained from a dataset with 40000 network flows, where 20000 are classified as malicious flows and the remaining as benign flows. To classify network flows as malicious or benign, three statistical tests of different natures were used: parametric (Pearson and Komolgorov correlation and p-value calculation) and non-parametric (Cramer-Von Mises p-value calculation), applied to the results of the frequency of occurrence of the first digit with the empirical frequency of each natural law.

Although the results obtained with the model based on the laws of Benford, Stigler and Zipf do not surpass the results obtained by the majority of models based on machine learning, as initial results, we emphasise that they are satisfactory, with a maximum F1 of 69.40% having been obtained.

Keywords: Network forensics, Benford's law, Stigler's law, Zipf's law, first digits law, statistical coefficient correlation,

1 Introduction

Illegitimate access to critical and confidential data motivates cybercriminals to exploit security flaws in a computer network's systems, whose illicit activity generates an unusual type of behaviour on the network, resulting in a series of attacks, including ransomware and phishing [2]. Performing a statistical analysis of network flows consists of analysing the metadata, making it possible to identify the machine(s) that have been compromised in a cyber-attack, highlighting this approach's portability and ease of handling data.

This work describes applying a set of statistical natural laws, namely Benford's, Stigler's and Zipf's Laws, as a model for analysing and detecting malicious flows in a computer network. The operation proposed by the present model is based on extracting the first digit from the characteristics existing in the public dataset CIC-IDS2017 [1], consisting of network flows obtained through the network traffic flow generator and analyser (CICFlowMeter). The dataset used in the experiments consists of 40000 network flows containing several types of attacks, namely, Brute Force (FTP and SSH), DoS, Web Attacks, Infiltration, Botnet and DDoS, obtained on different days of the week. Out of 40000 network flows, 20000 are benign flows, and 20000 are malicious flows, getting a balanced dataset.

To assess the robustness of the model, Pearson's and Komolgorov's coefficient correlations and the Cramer-Von Misses (CVM) goodness-of-fit test were calculated, which allowed the classification of network flows into malicious or benign.

2 Fundamentals of natural laws

Natural laws are usually identified as axioms and justified by mathematical rules and manipulations [3]. It is in this context that Benford's, Zipf's, and Stigler's Laws emerge, defined empirically as natural laws, and which allow us to conclude the existence of manipulations in data sets by the frequency with which each digit occurs (Benford's and Stigler's Law), or by the frequency of occurrence of words in texts (Zipf's Law).

Benford's law states that the frequency of occurrence of the digit 1 is 30.10%, of the digit 2 is 17.6% and so on, and the law can be extended to two or more digits. Theorem 1 mathematically defines the general Benford's Law [1, 4], and from this, the frequencies of occurrence of any digit can be calculated.

Theorem 1 (Benford's law) *Be* $k \in \mathbf{Z}$, $d_1 \in \{1,2,3,...,9\}$ *and* $d_j \in \{0,1,2,...,9\}$, $j = 2,...,k$.

$$P(D_k = d_k) = \log\left(1 + \frac{1}{\sum_{i=1}^{k} d_i \times 10^{k-i}}\right) \tag{1}$$

In 1945, Stigler proposed an alternative to the distribution of digits identified by Benford, in which he argued that the digits with the largest entry in the statistical table have the highest probability of starting with $d = 1,2,...,9$ and that the remaining entries in the table are obtained randomly from the uniform distribution of the smallest digits. Equation 2 allows us to determine the average frequency with which the first digits occur. Let $d \in \{1,2,3,...,9\}$,

$$P(d) = \frac{d\ln((d) - (d+1)\ln(d+1) + \left(1 + \frac{10}{9}\ln(10)\right)}{9} \tag{2}$$

In 1940, George Zipf defined the law on the frequency of occurrence of words in linguistic terms, where he established that the frequency of occurrence of any word is inversely proportional to its position in the frequency table. Later, Zipf's Law was extended to the occurrence of digits, where a weight was assigned to each digit according to its frequency of occurrence, i.e. if digit 1 occurs more frequently, it has a lower weight than the digit that occurs more frequently, bringing Zipf's Law closer to Benford's Law. The frequency of each digit is given by Equation 3.

$$F(r) = \frac{C}{r^{\alpha}} \tag{3}$$

where $F(r)$ represents the frequency of occurrence of each digit, C is a normalizing constant, r is the frequency rank of the digit, and $\alpha \approx 1$.

3 Model architecture

Figure 1 represents the general architecture of the model based on the individual application of the natural laws of Benford, Stigler and Zipf to identify malicious network flows. Since a public dataset was used, the process began with converting the files from `.csv` format to `.xlsx` format, as the procedure for obtaining the network flows had already been carried out previously. After converting the files, a set of network flows was extracted from each `.xlsx` file, representing benign and malicious flows, for a total of 40000 flows. A script built in Matlab was then applied to extract the first digit from the new data set. The main idea was to use the three natural laws independently of the first digit extracted from the characteristics of each flow and see if the model could identify which flows were considered malicious. Later, the total frequency of occurrence of all the digits was calculated based on all the network flows and the frequency of each digit for each network flow. This procedure aimed to check whether, on the one hand, there was a strong correlation between

[1] University of New Brunswick. Intrusion detection evaluation dataset, 2017.

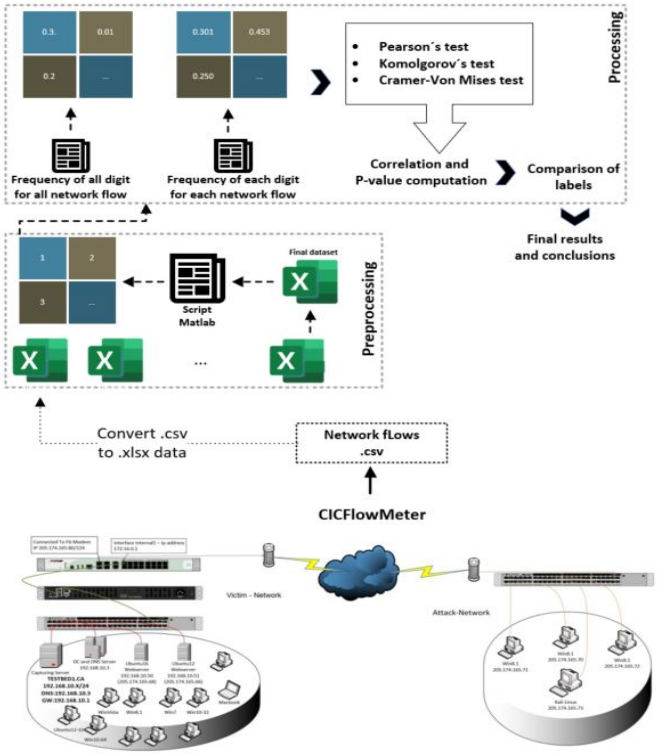

Figure 1: General architecture of the model based on the natural laws of Benford, Stigler and Zipf.

the total frequencies obtained and the empirical frequencies of each natural law and, on the other hand, to analyse flow by flow whether this correlation remained strong and whether it could serve as a decision-making factor in identifying malicious flows.

The decision was made by applying statistical inference, namely the Pearson, Komolgorov and Cramer-Von Mises tests, where the correlations and p-values between the frequencies of occurrence of each flow and the empirical frequencies were calculated. This procedure enabled the generation of a set of labels for each flow analysed, allowing these new labels to be compared with the original labels and thus classifying each network flow as malicious or benign.

Given the number of zeros in the dataset, not using them could result in losing important information and the model being misclassified as benign or malicious flows. It is common for attackers to use the zero digit to represent the absence of a value, allowing them to exploit vulnerabilities in computer systems and carry out DDoS attacks or other types of attacks. It should be noted that the empirical frequencies of natural laws start at the 1 digit and end at the 9 digit. As an initial solution, it was assumed that the digits of Benford's and Stigler's Laws did not end in the digit nine but in the digit ten. Thus, when counting the digits in the dataset, the script produced in MatLab assumes that there are ten occurrences, i.e. 1...10, where 1 represents the number of zeros and 10 represents the number of 9s in the dataset.

4 Results of the proposed model

Observing Table 1, the best result was obtained by applying Zipf's Law using Pearson's test, with a maximum F1 score of 68.98% and a recall of 92.05% for a significance level of 0.01. This result aligns with the application of Benford's and Stigler's Laws when using the same Pearson test with the same significance level of 0.01. By analysing Table 1, we can identify a pattern in the results, where regardless of the Natural Law used, the best results were obtained when a confidence level of 0.01 was used, then 0.05 and finally 0.1. Another important conclusion is the ineffectiveness of the Cramer-Von Mises test, with a high false negative rate, making the model ineffective in detecting malicious flows.

The number of true positives is high when ten digits are used, reflecting the zero digit's importance in this research. There is a high false positive rate, justified by the frequency of occurrences in the first four digits. These frequencies influence the results of the flow classifications, resulting in a correct classification if the frequencies are slightly higher than or identical to the empirical frequencies of the natural laws and an incorrect classification if the frequencies of the first digits are too high or too low concerning the empirical frequencies.

Several solutions can be used to minimise the number of false positives and increase the model's efficiency, such as using distance functions or creating an ensemble between classifiers.

Table 1: Overall results by individually applying Benford's, Stigler's and Zipfs' Laws when using ten digits. Legend: TP - True Positives ,TN - True Negatives ,FP - False Positives FN - False Negatives ,PR - Precision ,RE - Recall ,F1 - F1-Score AC - Accuracy.

Benford's Law	Degrees of significance	TP	TN	FP	FN	PR	RE	F1	AC
Pearson	0.05	20000	2	19998	0	0.500	1	0.6666	0.5000
	0.01	19998	610	19930	2	0.5077	0.9999	0.6735	0.5152
	0.1	20000	0	20000	0	0.5000	1	0.6666	0.500
Komolgorov	0.05	18642	1106	18894	1358	0.4966	0.9231	0.6480	0.4937
	0.01	19992	12	19988	8	0.5001	0.9996	0.6666	0.5001
	0.1	16740	1353	18647	3260	0.4731	0.8370	0.6045	0.4523
Cramer-Von Mises	0.05	3945	11415	8585	16055	0.3148	0.1973	0.2425	0.3840
	0.01	1116	17956	2044	18884	0.3532	0.0558	0.0964	0.4768
	0.1	6876	7677	12323	13124	0.3581	0.3438	0.3508	0.3638
Stigler's Law	Degrees of significance	TP	TN	FP	FN	PR	RE	F1	AC
Pearson	0.05	20000	466	19534	0	0.5058	1	0.6718	0.5116
	0.01	16468	4444	15556	3532	0.5142	0.8234	0.6331	0.5228
	0.1	20000	30	19970	0	0.5003	1	0.6669	0.5007
Komolgorov	0.05	15810	1381	18619	4190	0.4592	0.7905	0.5809	0.4298
	0.01	18840	527	19473	1160	0.4917	0.9420	0.6462	0.4842
	0.1	15759	1761	18239	4241	0.4635	0.7880	0.5837	0.4380
Cramer-Von Mises	0.05	2688	12637	7363	17312	0.2674	0.1344	0.1789	0.3831
	0.01	155	19178	822	19845	0.1586	0.0077	0.0148	0.4833
	0.1	5670	8666	11334	14330	0.3335	0.2835	0.3065	0.3584
Zipf's law	Degrees of significance	TP	TN	FP	FN	PR	RE	F1	AC
Pearson	0.05	19916	758	19242	84	0.5086	0.9958	0.6733	0.5169
	0.01	18411	5028	14972	1589	0.5515	0.9205	0.6898	0.5860
	0.1	20000	54	19946	0	0.5006	1	0.6671	0.5013
Komolgorov	0.05	15810	1381	18619	4190	0.4592	0.7905	0.5809	0.4298
	0.01	18840	527	19473	1160	0.4917	0.9420	0.6462	0.4842
	0.1	15759	1761	18239	4241	0.4635	0.7880	0.5837	0.4380
Cramer-Von Mises	0.05	2688	12637	7363	17312	0.2674	0.1344	0.1789	0.3831
	0.01	155	19178	822	19845	0.1586	0.0077	0.0148	0.4833
	0.1	5670	8666	11334	14330	0.3335	0.2835	0.3065	0.3584

5 Conclusion

This paper describes the application of natural laws to a dataset containing malicious and benign network flows to detect intrusions in a computer network. By calculating the correlations and p-values between the frequencies of occurrence of each digit obtained from the network flows and the empirical frequencies of Benford's, Stigler's and Zipf's laws, the model-based individually on the natural laws made it possible to classify the flows obtained on different days with different types of attacks. The first results from the research suggest the reliability of the Pearson test as a classifier, followed by the Komolgorov test and finally the Cramer-Von Mises test, where, according to Table 1, it achieved the worst results. The less good results obtained using natural laws can be explained by the few features in the data set, where, despite the high number of network flows, the number of features extracted is low, settling at almost 80 features. In future research, we plan to use 500 or more features. The high false negative detection rate combined with the model's speed is a good indicator that, together with the future resolution of the number of false positives, will make the model more efficient. By comparing the natural laws under investigation and observing Table 1, we can conclude that the model based on the natural laws produces similar results, both in terms of the classifier and the degrees of significance, so it will be necessary in future work to identify new classifiers, allowing us to create a robust model capable of rigorously and accurately identifying malicious network flows.

References

[1] Arno Berger and Theodore P. Hill. The mathematics of Benford's law: a primer. *Statistical Methods & Applications*, 30(3):779–795, jun 2020. doi: 10.1007/s10260-020-00532-8.

[2] Marta Fuentes-García, José Camacho, and Gabriel Maciá-Fernández. Present and future of network security monitoring. *IEEE Access*, 9: 112744–112760, 2021. doi: 10.1109/ACCESS.2021.3067106.

[3] Norman Swartz. *The concept of Physical Law*. Cambridge University Press, 2nd edition, 2003. ISBN 0-9730084-2-3.

[4] Luohan Wang and Bo-Qiang Ma. A concise proof of Benford's law. *Fundamental Research*, 2023. ISSN 2667-3258. doi: https://doi.org/10.1016/j.fmre.2023.01.002.