An Approach for Detecting Anomalies by Assessing the Inter-Arrival Time of UDP Packets and Flows Using Benford's Law

Ali Naghash Asadi
School of Computer Engineering, Iran University of Science and Technology, Tehran, Iran
*Corresponding Author's E-mail: aliasadi@comp.iust.ac.ir

Abstract
In this paper, from the perspective of Benford's law the inter-arrival time of UDP in packet and flow levels, is investigated. Benford's law is an empirical law that describes the distribution of first digits in series of numbers in natural phenomena. We claim that Benford's law describes the inter-arrival time of UDP packets and flows in normal traffic of networks. As a result, any significant anomaly in UDP packets and flows including deliberate intrusions, unwanted errors or in general, network failures, can be identified by checking the first digit distribution inter-arrival time of UDP packets and flows. In a recent work, the relationship between Weibull distribution and Benford's law was studied. In another work, the compliance of the inter-arrival time of UDP packets and flows from Weibull distribution is presented. In this paper, we have proposed a method for using Benford's law for detecting anomalies in inter-arrival time of UDP packets and flows. The proposed method can detect the UDP Flood attack with high detection rate.

Keywords: Network security; anomaly detection; Benford’s law; Weibull distribution; UDP traffic.

I. Introduction
Network security is an important issue in today world. A variety of methods are devised to solve network security issues. One of these issues is network traffic analysis, which is needed to discover anomalies. For this purpose, there are various techniques and tools that analyze either the structure or the behavior of the network packets and flows to detect anomalies. Then, by the detection of anomalous behavior, appropriate actions may be taken. The structural analysis of network traffic or misuse detection techniques use some patterns or rules to detect anomalies. On the other hand, behavioral analysis or anomaly detection technique, first understand the network normal behavior and any deviation from it is considered as anomalies. In addition, there are at least two different levels for network traffic analysis: packet level or flow level. Each of these levels of analysis has some advantages and disadvantages. Each level provides the details of the information, which is used in accordance with the objectives and performance required for detection of anomalies. Two of the important features in network are the large volume of network traffic and online data. These features shows online and high-volume data processing. Therefore, after identifying patterns and normal behavior of network traffic, the devising of algorithms for the detection of anomalies and the deviations on huge amounts of data will be required. In this article, Benford’s law and anomaly detection methods are introduced. Then, Weibull distribution as the most important distribution for modeling the behavior of network traffic and its compliance with Benford’s law will be given. Then, the compliance of inter-arrival time of UDP packets and flows with Weibull distribution is shown. At the end using an experiment, Benford’s law is used to detect anomalies and its efficiency is shown.
II. Background and Related Work

A. Benford’s Law

The basic idea of this law is provided by Frank Benford in 1938. By observing different books in the library, he found that first pages of all books are more dirty than other pages (readers give more attention to first pages), and as closer we go to the final pages of the book pages dirt decreases. With further studies, obtained logarithmic relationship for this phenomenon and extended it for other varieties numbers in nature. According to Benford’s law, the frequency of $d$, as the first digit of each of the elements of a data set, as it is expected is not uniform, but strangely, there is a logarithmic relationship [1]:

$$P_d = \log \frac{d+1}{10}$$

Although the law seems strange at first, but numerous sources proved this law. In addition, the researchers have defined conditions for data set that is in compliance with the law or is not. Moreover, this law is applicable for the first two digits or first three digits of a data set.

The law is used in several occasions and for various disciplines. Among these are the plans for the detection of tax evasion noted by Nigrini [2]. Nigrini showed that many aspects of the financial accounts such as costs claims follow Benford’s law. He presented special statistical tests that could detect fraud or error of the data. In addition, Buyse et al., in 1999, proposed using Benford’s law for fraud detection, as well as inadvertent errors in clinical tests [3]. Another generalization of Benford’s law is given by Fu et al. in 2007, which deals with modeling the first digit distribution of JPEG coefficients under different compression of Q-factor [4]. Another program in 2009 showed how we could use Benford’s law to investigate problems in the socio-economic data collected in the surveys [5]. In addition, this program was used to detect anomalies in the electoral data. Also, Benford’s law used in the study of natural sciences such as earthquakes and greenhouse gas emissions.

B. Anomaly Detection Methods

Phoha in 2002 noted that anomaly detection is determined by identification of deviations from normal behavior of a process [6]. Anomalies are mainly caused due to an error or malicious activity. Hence, anomaly detection include a wide range of applications including intrusion detection. Intrusion detection refers to detection of malicious activity of a computer system [7]. Intrusion detection systems can be classified into three approaches of pattern recognition, anomaly detection, or a hybrid approach [8]. A pattern recognition system, checks traffic patterns or data to detect a malicious activity. While anomaly detection systems, compare their occurred activity with a normal activity. Hybrid approach uses a combination of approaches. The main advantage of pattern recognition is that known attacks can be low false alarm rate is detected. This systems requires a pattern for every attack that secure network against attacks. The advantage of anomaly detection systems is the ability to detect new and unknown attacks, but these systems often suffer from a high percentage of false alarms.

General approach to anomaly detection is to define a normal behavior and by observing the actual behavior of the system and comparing it with the normal behavior, any deviations from it should be considered as an anomaly. However, there are many challenges regarding this simple approach. The key point about the difficulty of defining normal behavior is that this behavior should include any possible natural behaviors. Because, the anomalies change in accordance with the time and the attackers change their attack based on intrusion detection systems and methods. Another major challenge is the massive amounts of data. Anomaly detection techniques need effective calculations to handle the large volume of incoming data. In addition, data are usually online and requires that these calculations and analyzes to be performed online. Another important issue that arises due to the high volume of input is that even a low percentage of false alarms can overwhelm the analysis [9].
C. Weibull Distribution and Benford’s Law

The random variable $T$ is a Weibull random variable, if the probability density function (PDF) is as follows [9]:

$$f_T(t; k, \lambda) = \frac{k}{\lambda} (\frac{t}{\lambda})^{k-1} e^{-\left(\frac{t}{\lambda}\right)^k} \quad t \geq 0, k, \lambda > 0$$

(2)

where, $k$ and $\lambda$ are respectively the shape and scale parameters of the distribution. In addition, the cumulative distribution function (CDF) of the Weibull random variable is defined as follows [9]:

$$F_T(t; k, \lambda) = 1 - e^{-\left(\frac{t}{\lambda}\right)^k} \quad t \geq 0, k, \lambda > 0$$

(3)

Another important point about Weibull distribution is that PDF for $k<1$ severely reduces, so that the $+\infty$ started at $t=0$ is zero at $t=5$. In addition, it severely reduces for $k=1$, so that the $1/\lambda$ at $t=0$, starts and with increase of $t$, rapidly it reduces to zero [9]. However, for $k>1$ it is non-uniform. Clearly, this is displayed in Fig.1.

In [9], it is experimentally proven that Weibull distribution random variable follows Benford’s law. Therefore, easily and without loss of generality of issue, the Weibull conformance testing can be replaced with the first digit test, which is less complex. For empirical proof of this issue real and positive subset of $E_d$, for $d=1, 2, \ldots, 9$ is defined in [9]:

$$E_d = \bigcup_{p=-\infty}^{+\infty} [d10^p, (d+1)10^p)$$

(4)

Accordingly, the probability that the first digit of the random variable $T$ equals $d$ is the following equation [9]:

$$P_d = \Pr(T \in E_d)$$

(5)

In the case of Weibull random variable, it can be formulated as follows [9]:

$$P_d(k, \lambda) = \sum_{p=-\infty}^{+\infty} (F_T((d+1)10^p) - F_T(d10^p))$$

(6)

Moreover, the above equation is equivalent to the following equation [9]:

$$P_d(k, \lambda) = \sum_{p=-\infty}^{+\infty} (F_T((d+1)10^p) - F_T(d10^p))$$
The probability for different values of the shape and scale parameters is estimated and is shown where about k<1, the probability is log (1+1/d) (according to Benford’s law). The Euclidean distance (ED) defined in equation (8) is used for calculation [9]:

\[ ED = \sum_{d=1}^{9} (\hat{p}_d - e_d)^2. \]  

In this equation, \( P_d \) and \( e_d \) are calculated by equations (7) and (1), respectively. The obtained ED results are shown in Fig. 2. It shows that at various parameters of the scale \( \lambda \) and shape \( k \), the error is very small.

**D. Follow Inter-Arrival Time of UDP Packets and Flows from Weibull Distribution**

In [10], it was shown that the inter-arrival time of UDP packets and flows complies with Weibull distribution. In Fig. 3, the distribution of inter-arrival time of UDP packets can be seen. In [10], the inter-arrival time of UDP packets and flows to match the mathematical distributions is studied by Kolmogorov-Smirnov method. Among the existing distributions, Weibull distribution, Pareto, gamma and lognormal have shape and scale parameters, but only exponential distribution include a rate parameter. Table.I, which is derived from this reference, indicates appropriateness of inter-arrival time of UDP packets with any of the distributions. This is shown with KS that is Kolmogorov-Smirnov criterion shows and the maximum difference. As a result, smaller the criterion, appropriate the distribution.

In this work, the inter-arrival time of UDP flows is modeled using the same approach. However, UDP, due to the similar nature of inter-arrival time of packets and flows of UDP packets, it seems that the proof of compliance of inter-arrival time of packets from a mathematical distribution represents time compliance of the flows of the same mathematical distribution.


<table>
<thead>
<tr>
<th>Distribution</th>
<th>Param. 1</th>
<th>Param. 2</th>
<th>KS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weibull</td>
<td>0.734</td>
<td>0.003</td>
<td>0.053</td>
</tr>
<tr>
<td>Pareto</td>
<td>1.848</td>
<td>0.004</td>
<td>0.046</td>
</tr>
<tr>
<td>Gamma</td>
<td>0.480</td>
<td>0.008</td>
<td>0.103</td>
</tr>
<tr>
<td>Exponential</td>
<td>249.330</td>
<td>-</td>
<td>0.176</td>
</tr>
<tr>
<td>Lognormal</td>
<td>1.545</td>
<td>-6.472</td>
<td>0.053</td>
</tr>
</tbody>
</table>

Also, in [10], the Pareto distribution as nearest distribution to the inter-arrival time of UDP packets, but it can be seen that the Weibull distribution with a very small distance is second and this distribution can be introduced as the simulation of inter-arrival time of UDP packets. The shape and scale parameters for the simulation of inter-arrival time of UDP packets at this reference (as shown in Table I) are assumed $k = 0.734$ and $\lambda = 0.003$, respectively.

### III. Experiments

#### A. Validation Methodology

We have conducted two experiments to support our claim about the relationship between inter-arrival time of UDP packets and flows with Benford’s law. The first experiment is performed on the inter-arrival time of UDP packets. In this experiment, each UDP packet individually participate in the experiment and we have considered the inter-arrival time of the packets. In the second experiment, the inter-arrival time of UDP flows are examined. To obtain flows of UDP, first the packets based on source and destination IP address and port are grouped together, so that the number of categories will be equal to the number of flows. Then, in each group, if the distance is more than a second to the next packet, the flow will be divided into two flows. In other words, the inter-arrival time of UDP flows is considered more than a second, because a source and destination IP and port at two different times may send packets to each other and their flows must be distinguished. After determining the flows, the lowest arrival time in each category is referred to as the arrival time of the flow. After sorting flows based on arrival time, the time difference of the arrival of each flow from the previous one is considered as the inter-arrival time of the UDP flows. These experiments consist of two phases. In the first phase, the frequency of the first three digits on the normal data set is calculated and the result is compared with expected logarithmic distribution by the application of appropriate measures. Sum-of-squared-error (SSE) is the appropriate measure, which will be described below. This experiment is conducted for the first digit and the first two digits of inter-arrival time of UDP packets and flows and there was no change in the results. But, because of the large number of first digits or first two digits of zero in the data set, and the impossibility of calculating the logarithmic distribution (according to equation (1)), it was decided to use the frequency of the first three digits.

In the second phase, attack packets and flows will be added to the normal data and changes in the frequency of the first three digits inter-arrival time of UDP packets and flows and the value of SSE will be examined. Thus, it will be shown that the anomaly packets and flows can affect relationships graph. Also, using Benford’s law one can identify anomalies as well as anomaly packets and flows.

#### B. Traffic data set

The traffic data set that is used in these experiments is the MAWI traffic, which is obtained from the backbone of the Pacific Ocean that connects the combination of hosts, public and academic services to the internet [11]. The MAWI traffic data set is a public data set extracted from the archived traffic of MAWI working group. We just used the part of the Pacific Ocean backbone with
the speed of 150Mbps. Traffic set file includes IP packets sent on this link, from 14 to 14:15 on 01/11/2014. To reduce the data volume, only a minute of traffic examined and UDP packets extracted from it. The final file includes UDP packets set for 14 to 14:01 from the main traffic set, which is 327,740 UDP packet. In addition, by the classification of UDP packets based on source and destination IP and port, 118,584 traffic flows in this collection were detected with the same the source and destination IP and port address. In order to detect anomalies, UDP Flood attack is added to normal traffic set. In this attack, different systems with various IP addresses and ports send a large number of packets on a special IP and port to engage that system and its port. Our attack was simulated by UDP Unicorn software [12] with a rate of 10 packets per second during 300 seconds on a specific port and IP.

C. The Method of Experiment

To perform these experiments, the inter-arrival time of UDP packets/flows of the data set is obtained by the subtraction of arrival time of packets/flows from the previous arrival time of packets/flows. Due to the small amount of inter-arrival time, these values were multiplied in a larger number to derive at least three integer numbers. Integer values obtained are considered as the first three digits inter-arrival time of UDP packets/flows. Then, the frequency of each of the values 1 to 999 is counted and shown in the graph. This graph shows logarithmic nature of the first three digits frequency inter-arrival time of UDP packets/flows. The SSE difference criterion was used for a better understanding. It is calculated by the following equation:

\[ SSE = \left( \frac{No(d)}{part} - \log_{10}(d+1)d \right)^2 \]  

In this equation, No(d) is the number of packets/flows that their first three digits is equal to d. Also, part is the number of the packets/flows that were studied. Calculations is done in certain periods for greater precision, for this purpose, the variable part is used. For example, if there were 100,000 packets/flows, first No(d) in the 20,000 packets/flows is calculated to obtain the SSE, then, this calculation is repeated for the second 20,000 packets/flows. This helps better determining the location of anomalies.

The part size determination is very important in the diagnosis of disorder. The part size must not be large enough to hide anomalies, and not so small that logarithmic equation of frequency of the first digits cannot be done. This amount will be estimated through several tests. After determining the amount of part and the doing the calculations, SSE of each packet/flow will be shown in the graph. Since this graph is the result of comparison of frequencies of the first 3 digits of the inter-arrival time of normal UDP packets/flows with logarithmic distribution, small SSE is created by any packet/flow. Then, to demonstrate the efficacy of using Benford’s law in anomaly detection, packet arrival times of UDP Flood attack will be added at a specified location of the data set of arrival time of normal UDP packets. The above process will be repeated at the same way, again. First 3 digits frequency and achieved SSE graphs show differences. In SSE graph, it can be seen that at the same places of UDP Flood attack packets, SSE is a relatively high which is different with other values of normal packets SSE.

IV. Results of Experiments

Four graphs are shown for each experiment. The first two graphs are related to sets of normal traffic and second two graphs are related to a set with traffic anomalies. Above graphs, in any figure, shows the frequency of inter-arrival time for the first three digits. The horizontal axis is the first digits of 1 to 999 and the vertical axis is the frequency of each digit. Below graphs in each figure shows calculated anomalies for each packet. The vertical axis shows the difference and the horizontal axis is the number of each packet. The graphs related to the set of normal traffic, logarithmic nature of the frequency of first 3 digits of the inter-arrival time is shown. It is shown that there is no deviation in the packets. In the graphs related to set of anomalies traffic, logarithmic nature of frequency of first 3 digits inter-arrival time is violated. Anomalies were observed in some packets.
A. In Inter-arrival time of packets

MAWI traffic set includes 327,740 UDP packets. Part size of this study is determined to be 20 thousand, which forms 17 groups of packets (Fig.4). Then, 3000 arrival time of UDP Flood attack is at 200 thousand packet of normal data set. The resulting data set will be equal to 330,740 arrival time of UDP packet. Then, experiment in the same way is done on new data set. In this experiment, 17 groups of packets are created. Fig.5 shows the results of these tests. It is observed that in packets after 200 thousand, SSE has increased a lot. This is due to the change in the frequency of the first 3 digits of normal data set.

B. In Inter-arrival time of flows

MAWI traffic set includes 118,584 UDP flows. Part size of this study is determined to be 20 thousand, which forms 7 groups of flows (Fig.6). Then, 3000 arrival time of UDP Flood attack is at 200 thousand packet of normal data set. The resulting data set will be equal to 121,584 arrival time of UDP flow. Then, experiment in the same way is done on new data set. In this experiment, 7 groups of packets are created. Fig.7 shows the results of these tests. It is observed that in flow after 70 thousand, SSE has increased a lot. This is due to the change in the frequency of the first 3 digits of normal data set.

V. Evaluation

In this section, anomaly detection method proposed by Benford’s law will be examined more closely. The criteria considered in this section are as follows: Time of detection of anomalies, the percentage of false positive error and percentage of detection. Percentage of false positive error shows normal packets/flows that mistakenly have been detected as attack. In addition, percentage of detection is packets/flows that are properly known as attack. Table.II and Table.III show a summary of the results. As noted above, the range of packets/flows of study (part) plays an important role in the diagnosis of anomalies in Benford’s Law.
To detect anomalies in the packets/flows arrival time are examined at ranges of 5,000, 20,000 and 40,000. As can be seen with the increase of the range the number of groups of packets/flows decreases, but this has no direct impact on the detection rate and false positive error percentage and time required to calculate the deviation. Determining the range of packets/flows need to be done carefully in order to not to have large enough ranges to let anomalies to be hidden and not to have so small ranges that logarithmic equation could not be set. In all the results, it can be seen that with the right choice of packet/flow error detection rate and false positive error percentage, is very appropriate and this means that Benford’s law is efficient in network traffic anomaly detection.

![Fig. 6. Frequency of the first 3 digits of inter-arrival time of flows and the amount of SSE calculated in normal mode](image1)

![Fig. 7. Frequency of the first 3 digits of inter-arrival time of flows and the amount of SSE calculated in anomaly mode](image2)

<table>
<thead>
<tr>
<th>TABLE II. INTER-ARRIVAL TIME OF UDP PACKETS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Part</td>
</tr>
<tr>
<td>------</td>
</tr>
<tr>
<td>5000</td>
</tr>
<tr>
<td>20000</td>
</tr>
<tr>
<td>40000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TABLE III. INTER-ARRIVAL TIME OF UDP FLOWS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Part</td>
</tr>
<tr>
<td>------</td>
</tr>
<tr>
<td>5000</td>
</tr>
<tr>
<td>20000</td>
</tr>
<tr>
<td>40000</td>
</tr>
</tbody>
</table>

**Journal of Knowledge-Based Engineering and Innovation (JKBEI)**
Universal Scientific Organization, [http://www.aeuso.org/jkbei](http://www.aeuso.org/jkbei)
ISSN: 2413-6794 (Online)
Conclusion

In this article, we have used Benford's law as a criterion to detect anomalies at inter-arrival time of UDP packets/flows. The reason for using Benford's law is due to its fast performance in detection of anomaly against other criteria. Only by assessing the first digits of inter-arrival time of packets/flows, anomalies could be identified. In two previous works, the compliance of Weibull distribution to Benford's law and inter-arrival time of UDP packets and flows compliance to the Weibull distribution have been studied. In this paper, by knowing these two facts and combining them with each other, we introduced Benford's law as a fast and valuable criteria for early detection of anomalies of the inter-arrival time of UDP packets and flows. Therefore, it is no longer necessary to check if the inter-arrival time of UDP packets or flows is Weibull distribution. However, anomalies can be detected with Benford's law just by looking at the first digits of their inter-arrival time and there is no need to set the parameters.

References