DOI: https://doi.org/10.31410/ITEMA.2023.11

Linear Statistical Features for the Purposes of Computer Network Automation

Milan Milivojević[1]
Milan Pavlović[2]
Marija Zajeganović[3]

Received: December 22, 2023
Accepted: March 20, 2024
Published: May 28, 2024

Keywords:
Computer networks;
Automation;
Python;
Statistical analysis

Abstract: *The development of artificial intelligence finds applications in various practical areas, such as computer networks. The primary goal of introducing automated methods is the efficient optimization of computer network operations. As input data for each algorithm, parameters that best describe the computer network are defined. Therefore, it is essential to define which parameters are relevant to ensure the significant use of automation methods. Parameters like Round-Trip Time delay resulting from ping commands can be used as the basis for defining input parameters for the automation system. This approach can help detect anomalies in network operation due to topology disruptions and increased load on specific links within the network. These anomalies can be mitigated by adjusting routing protocol parameters and activating redundant links. The authors describe basic features that can be extracted from time series data containing information about delay times. Special emphasis is placed on the characteristics that result from linear statistical analysis using the Python programming language.*

1. INTRODUCTION

In recent times, there has been evident development of artificial intelligence tools in various spheres of human activity. It is not necessary to further emphasize the significance of artificial intelligence in terms of increasing the efficiency of various processes and systems while reducing costs, ensuring their further development. In the field of computer networks, automation is increasingly prevalent. Computer networks are complex systems consisting of various elements. Primarily present are network elements such as routers and switches, which have their predefined configurations. Since the conditions in the operation of computer networks are not static, it is necessary to make occasional changes in the configuration to ensure smooth operation. Computer networks are susceptible to faults in the elements themselves, congestion on certain links within the network, as well as external links. The intensity of disturbances can vary from minor faults to failures of individual parts of the network. On the other hand, malicious attacks can also cause problems in the network's operation. In this case, it is necessary to make certain modifications within the network to successfully carry out the defense and return to normalcy with minimal degradation (Yaibuates & Chaisricharoen, 2020).

The time to restore normal functioning must be as short as possible. The automation system can react by disabling certain parts, forming appropriate isolated area, disconnecting specific links, creating new routes, etc. To make all of the above possible, it is necessary to automatically change configurations, primarily of network elements. In the case of routers, this may

[1] Academy of Technical and Art Applied Studies Belgrade (ATUSS) – Department ICT College for vocational studies, Zdravka Čelara, 16, 11000, Belgrade, Republic of Serbia
[2] Academy of Technical and Art Applied Studies Belgrade (ATUSS) – Department ICT College for vocational studies, Zdravka Čelara, 16, 11000, Belgrade, Republic of Serbia
[3] Academy of Technical and Art Applied Studies Belgrade (ATUSS) – Department ICT College for vocational studies, Zdravka Čelara, 16, 11000, Belgrade, Republic of Serbia

involve changes to interface configurations, modifications to routing protocol parameters, adjustments to the costs of individual links, and traffic balancing to redirect traffic along alternative paths and methods.

The automation system operates with specific scripts that gather all the necessary information for the implementation of the automation procedure. First and foremost, it is essential to collect information about the network's status, and this process occurs in cycles. After gathering all the necessary information, the computer network automation system optimizes the network, undertaking urgent procedures in the case of unforeseen situations (Ahuja et al., 2021).

This paper presents linear statistical characteristics that can serve as a starting point when the automation system assesses the performance of a computer network. The advantage of applying these features lies in their simplicity of calculation. By using the well-known 'ping' command, information about the availability of specific hosts in the network and latency time can be obtained. Basic statistical analysis was then performed on this data to extract certain features, in this case, moments up to the fourth order. Additionally, the possibility of a more in-depth analysis using the linear regression procedure was demonstrated. Based on these extracted features, the automation system can make certain modifications to the computer network to ensure smooth operation. The Python programming language was used to implement specific simulations.

The paper is organized into five chapters. The first chapter serves as an introduction. The second and third chapters define the Round-Trip Time (RTT) and the acquisition of that time. The fourth and fifth chapters are dedicated to linear statistical analysis and modeling of the RTT time series. The sixth chapter provides a conclusion.

2. ROUND TRIP TIME

The Round-Trip Time (RTT) represents a fundamental parameter that quantitatively describes the performance of networks in general, particularly in the context of computer networks. The content transmitted through computer networks can vary significantly in its characteristics, making the RTT parameter directly related to user experience. In the case of video content streaming, it is directly correlated with the quality of the user's video experience (video QoE). On the other hand, it directly affects the page load time of a website. Any change in RTT statistics indicates a degradation in user experience. The cause of this degradation primarily lies in issues within the computer network itself, such as problems with network elements of hardware and/or software nature, inappropriate network protocol settings, interruptions in individual links, etc. Secondly, problems may arise from malicious activities, such as various types of network attacks aiming to intercept traffic and ultimately make the service unavailable. Therefore, sudden and significant variations in the RTT parameter indicate issues such as network congestion or that the network is under attack from remote entities. Hence, monitoring the RTT parameter is of great importance for overseeing a computer network to react promptly and eliminate adverse effects within the shortest possible time frame.

The most commonly used tool for evaluating the Round-Trip Time (RTT) parameter is the application of ICMP (Internet Control Message Protocol), which is based on the exchange of request/reply messages between two nodes in the network. Nodes in the network refer to any host (computer, corresponding serial/Ethernet interface of routers or switches). The ICMP

protocol, in synergy with the IP protocol, provides the capability to report potential errors that occurred during the transmission of IP packets on the path from the source computer to any host (Gezer, 2019).

There are various types of ICMP, but the use of the ping option is crucial for us. Ping is another term for sending ICMP Echo messages. In this case, the header size is 8 bytes, where the corresponding type field is used to differentiate between requests and response messages. The value 8 represents a request (Echo request message), while 0 is reserved for a response (Echo reply message). This is illustrated in Figure 1.

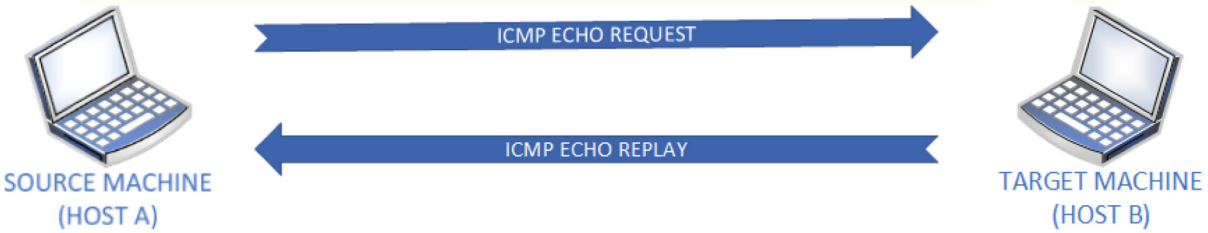

Figure 1. Ping through ICMP messages

Source: Own processing

In general, two points are observed (source and destination). The first point is usually within the computer network, while the second point can be a remote node outside the boundaries of the observed network. The connections between nodes represent links. Links can be direct, or there may be hidden routers serving as links between individual segments that are not explicitly visible. When measuring RTT, it is assumed that there is no change in topology, and the routing table has an unchanged content.

The evaluation of RTT time is based on the exchange of ICMP request/echo messages between two nodes. The result of the operation can be that the host in node B is unreachable, or an echo message containing information about the RTT time, usually expressed in milliseconds (ms).

The Round-Trip Time (RTT) represents the total time required for a message to propagate from node A to node B and back. This time includes all delays originating from the link, processing, and waiting times. In this way, it is possible to answer whether the host is available and what time is required for packets to propagate to the host and back. If the host is unavailable, then the value of this parameter is infinite.

One drawback of estimating the RTT parameter based on ICMP requests is that Internet Service Providers (ISPs) often block or limit the flow of ICMP traffic. ICMP traffic has lower priority on routers (Guo & Heidemann, 2018). However, as a fundamental measure of network connectivity, it remains the primary choice for estimating RTT traffic. If certain problems are detected, more precise methods for estimating RTT time, available in the literature, can be considered in a second iteration (Sengupta et al., 2022).

3. ACQUISITION OF ROUND-TRIP TIME DATA

The process of acquiring the Round-Trip Time (RTT) data typically involves sending a series of ICMP requests to a specific node, measuring the time between sending the request and receiving the response for each ICMP request and recording these values in an array. These data can

then be analyzed to gain insights into the performance characteristics of the network, as well as to monitor changes over time (Matcharashvili et al., 2020).

During the acquisition of the RTT data, it is important to consider factors such as network conditions, network load, and potential communication issues with the remote node. This information can be crucial for maintaining and optimizing the performance of a computer network.

The core part of the Python script consists of the ping command from the pythonping Python library, which takes the following input parameters: the destination IP address, packet size in bytes, and the number of packets sent to the destination. For the destination host's IP address, you can choose a host belonging to the same local computer network as the source host or a host belonging to a different remote computer network. A packet size of 32 bytes and a packet count of 200 have been chosen empirically, and these values directly affect the time required to collect the resulting RTT array. Therefore, the selection of these values is crucial.

As a result of the ping command, data is printed, and it needs to be parsed to extract only the numeric data related to the RTT parameter expressed in milliseconds. For this purpose, a specific script has been written in the Python programming language.

Computer Code

```
import numpy as np
from pythonping import ping
def prikupi _ ping(ip _ adresa,broj,velicina):
    odgovori = ping(ip _ adresa,verbose=False,count=broj,size=velicina)
    RTT=[]
    for odgovor in odgovori. _ responses:
        x = str(odgovor)
        x = x.partition('in ')[-1][0:-2]
        RTT.append(x)
    RTT = np.array(RTT, dtype = float)
    return RTT
```

The extracted array with RTT data is stored in a separate file suitable for working with data processing programs and the Python programming language. In case the destination host is unreachable or for some reason, it is not possible to estimate the delay time, a non-numeric NaN-type value is stored in the file. The total number of values must correspond to the number of sent packets. In Figure 2, the dependence of the delay time expressed in milliseconds on the sequential number of sent packets is shown. It is observed that there is a pronounced explosiveness in the response, considering the occurrence of pronounced local maxima and minima.

In practical applications, it is often convenient to observe time series with certain averaging within a defined time frame. For this reason, moving averaging is used according to formula 1.1.

$$RTT^{*}[k] = \frac{1}{W}\sum_{n=k}^{W+k-1} RTT[n] \tag{1.1}$$

In this equation, W represents the length of the time frame over which averaging is performed, RTT is the original array with estimated delays, and RTT* is the averaged array with estimated delays.

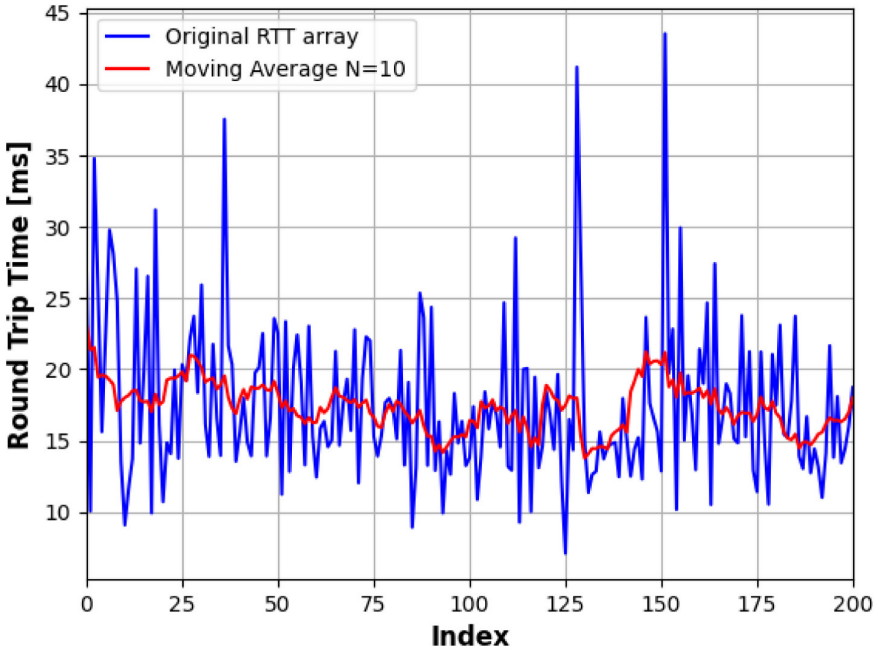

Figure 2. The original and moving averaged RTT time series

Source: Own calculations

The sample value at time k is obtained as the mean of the W previous samples, corresponding to the length of the time frame. For small values of this parameter, information about changes in delay values near the observed moment is retained. For larger values of the time frame length, information about local changes is lost in favor of a global picture of delay behavior. Figure 2 illustrates the averaged data (red line) alongside the original data (blue line) with a time frame length value of W=10. Explosive changes in delay values are no longer visible, while information about the data trend is preserved. The automation system for computer networks controlled by artificial intelligence, through variations in the time frame length parameter, influences the nature of the required delay data in the network. Any irregularity in the operation of the computer network will directly reflect in the appearance of the time series and can be easily detected.

4. LINEAR STATISTICAL ANALYSIS OF THE RTT ARRAY

Statistical analysis of Round-Trip Time (RTT) delays is crucial for analyzing and testing the functionality of computer networks. A broad categorization of statistical analyses is based on the use of linear or nonlinear methods and models on the time series. Basic analysis involves the application of linear methods, with the linear regression model being the most commonly used. Among the fundamental parameters calculated for each time series, moments up to the fourth order are typically employed, including mean, standard deviation, skewness and kurtosis (Gupta & Kapoor, 2018).

The mean, or the first-order moment, represents the arithmetic average of all values within a set or time series. All weights assigned to the values within the set have equal value.

$$\overline{RTT} = \frac{1}{N}\sum_{n=1}^{N} RTT[n] \tag{1.2}$$

The standard deviation, or the second-order moment of a time series, represents the square root of the arithmetic mean of the squared deviations of values in the series from their mean. The square in the formula for calculating the standard deviation ignores the sign of the deviations.

Standard deviation is one of the measures of data dispersion and is highly sensitive to the presence of extreme values within the series.

$$\sigma(RTT)=\sqrt{\frac{1}{N-1}\sum_{n=1}^{N}\left(RTT[n]-\overline{RTT}\right)^2} \quad (1.3)$$

Skewness is the third-order moment and represents the lack of symmetry in the distribution of time series values. An example of a symmetric distribution is the normal or Gaussian distribution. The sign of the third-order moment indicates asymmetry, which is pronounced on the right or left side of the distribution.

$$s=\frac{1}{N-1}\cdot\frac{1}{\sigma^3(RTT)}\sum_{n=1}^{N}\left(RTT[n]-\overline{RTT}\right)^3 \quad (1.5)$$

Knowing measures of central tendency such as skewness and dispersion still may not provide a complete understanding of the behavior of a time series distribution. Pearson introduced the concept of kurtosis (convexity of a curve), or the fourth-order moment. Kurtosis enables us to understand the flatness or peakedness of the curve. A curve flatter than the normal curve is known as platykurtic, while a curve more peaked than the normal curve is called leptokurtic.

$$k(RTT)=\frac{\frac{1}{N}\sum_{n=1}^{N}\left(RTT[n]-\overline{RTT}\right)^4}{\sigma^4(RTT)} \quad (1.6)$$

The discrete distribution or relative histogram of the observed RTT array has been estimated for graphical representation. In Figure 3, the relative histogram is displayed, with the bin frequency value set to 0.5 ms. The relative histogram is fitted using a Gaussian distribution, indicated by the red color. The parameters of the fitted Gaussian distribution correspond to the first and second-order moments. The values of all calculated moments are also shown on the same graph.

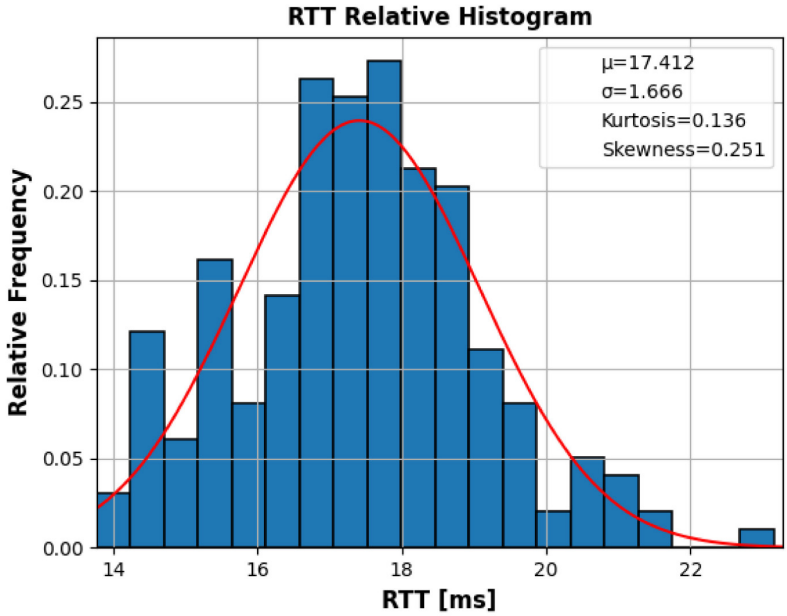

Figure 3. The relative and fitted histogram for RTT time series

Source: Own calculations

The relative histogram together with the moments provides a fundamental set of information about the state of the computer network when observing two nodes in the network. This information

must be periodically collected in the computer network automation system to prevent or effectively mitigate the negative effects during incident situations, such as various types of faults and failures. By observing two measurements of these parameters, it is possible to define specific thresholds that indicate which changes in parameter values may indicate certain performance degradations or impending attacks within the computer network. This allows for a more comprehensive investigation into the nature of the degradation when such changes are detected.

5. LINEAR MODELING OF RTT TIME SERIES

It has already been mentioned that the ping command used to assess RTT delay values takes multiple input parameters, such as the number of sent packets (count) and the size of the packets. In order to better analyze the nature of the RTT time series, a simulation was performed by varying one of the parameters. The simulation was conducted in multiple iterations, where each iteration involved an increase in the number of sent packets in the range from 40 to 100 with a step size of 5. For each value of sent packets, the corresponding RTT array was obtained. The first moment value was calculated for each of these RTT arrays. This process resulted in a set of points representing ordered pairs of the number of sent packets and the mean value of the RTT array. Linear regression was then applied to this set of points, obtaining parameters corresponding to the slope and intercept. These parameters define the line that fits the data and is shown in Figure 4.

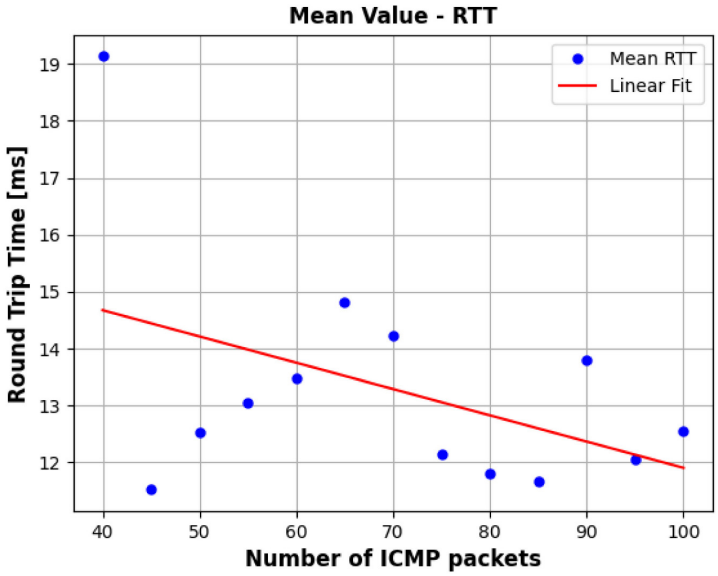

Figure 4. An example of linear fit for mean RTT values

Source: Own calculations

The parameter values of the linear model can be calculated based on the following equations (Lauritzen, 2023):

$$k = \frac{\sum_{i=1}^{N}\left(X[i]-\overline{X}\right)\cdot\left(Y[i]-\overline{Y}\right)}{\sum_{i=1}^{N}\left(X[i]-\overline{X}\right)^{2}}, n = \overline{Y} - k\cdot\overline{X} \tag{1.7}$$

where: $\overline{Y}$ is the mean of the dependent variable (mean RTT), $\overline{X}$ is the mean of the independent variable (mean number of sent packets), X_i and Y_i are individual data points and N is the number of data points. Calculated values for the slope coefficient and the intercept are: k= -0.044 and n=16.4.

Changes in these parameters depending on the moment when a specific test is conducted in the network serve as an additional indicator that can aid in network performance analysis. Any significant change indicates a potential issue in the network operations.

6. CONCLUSION

Automation in computer networks brings numerous benefits and facilitates administration, optimization, and configuration. All of this aims at detecting and resolving issues that may arise due to various external and internal factors. A prerequisite for any automation is defining a set of features that best reflect the state of the computer network. By activating specific scripts, data on the values of these features is collected. Based on the gathered features, the automation system, which may be controlled by artificial intelligence in the future, makes decisions about changing the configuration of the computer network in conditions that correspond to potential incident situations. This paper explores the possibilities of using linear statistical features for the automation of computer networks. The used features include moments up to the fourth order. The use of a linear regression model for automation purposes is demonstrated. For further research, it is necessary to apply nonlinear statistical approaches and methods. Linear statistical features are significant for rough examinations, while the application of nonlinear models is inevitable for a more detailed analysis.

References

Ahuja, N., Singal, G., Mukhopadhyay, D., & Kumar, N. (2021). Automated DDOS attack detection in software defined networking. *Journal of Network and Computer Applications, 187*, 103108. https://doi.org/10.1016/j.jnca.2021.103108

Gezer, A. (2019). Large-scale round-trip delay time analysis of IPv4 hosts around the globe. *Turkish Journal of Electrical Engineering and Computer Sciences*: Vol. 27: No. 3, Article 31. (pp. 1998-2009). https://doi.org/10.3906/elk-1803-137

Guo, H., & Heidemann, J. (2018). Detecting ICMP Rate Limiting in the Internet. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 10771 *LNCS*, 3–17. https://doi.org/10.1007/978-3-319-76481-8_1

Gupta, S. C., & Kapoor, V. K. (2018). Fundamentals of mathematical statistics (11th edn. (Thoroughly revised). Sultan Chand & Sons.

Lauritzen, S. (2023). *Fundamentals of Mathematical Statistics (I st ed.).* New York, United States: Chapman and Hall/CRC.

Matcharashvili, T., Prangishvili, A., Tsveraidze, Z., & Laliashvili, L. (2020). Scale Features of a Network Echo Mechanism: Case Study for the Different Internet Paths. *Journal of Computer Networks and Communications, 2020*, 1-9. https://doi.org/10.1155/2020/4065048

Sengupta, S., Kim, H., & Rexford, J. (2022). Continuous in-network round-Trip time monitoring. *SIGCOMM 2022 - Proceedings of the ACM SIGCOMM 2022 Conference* (pp. 473–485). https://doi.org/10.1145/3544216.3544222

Yaibuates, M., & Chaisricharoen, R. (2020). A Combination of ICMP and ARP for DHCP Malicious Attack Identification. 2020 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT & NCON) (pp. 15-19). 10.1109/ECTIDAMTNCON48261.2020.9090760