Network traffic classification based on periodic behavior detection

Josef Koumar
CTU FIT & CESNET a.l.e.
Prague, Czech republic
koumajos@fit.cvut.cz

Tomáš Čejka
CTU FIT & CESNET a.l.e.
Prague, Czech republic
cejkat@cesnet.cz

***Abstract*—Even though encryption hides the content of communication from network monitoring and security systems, this paper shows a feasible way to retrieve useful information about the observed traffic. The paper deals with detection of periodic behavioral patterns of the communication that can be detected using time series created from network traffic by autocorrelation function and Lomb-Scargle periodogram. The revealed characteristics of the periodic behavior can be further exploited to recognize particular applications. We have experimented with the created dataset of 61 classes, and trained a machine learning classifier based on XGBoost that performed the best in our experiments, reaching 90% F1-score.**

***Index Terms*—periodic commuynication, time series, periodicity detection, Lomb-Scargle periodogram, classification by periodicity, network traffic analysis, encrypted network traffic**

I. Introduction

Monitoring systems provide useful information for network operators and security teams. However, with the rise of encrypted communication, monitoring tools as well as detection systems lose visibility. There are various initiatives to increase the privacy of users on the internet and protect them from tracking or fingerprinting, such as TLS 1.3 encrypts more headers, and there is a draft proposing to encrypt even Server Name Indication (SNI) extension. Therefore, it is necessary to research new methods to recognize encrypted network traffic to maintain situational awareness of network operators.

Network traffic is usually analyzed using time series, which is a well-known approach in various domains such as statistics, economy, physics, and so on. However, network traffic in general is very hard to predict or model and many activities by users are nondeterministic. Contrary, there are many regular connections in the background that are promising for automatic detection, and thus they are the main concern of this paper.

This paper deals with network traffic analysis based on time series computation and measurement. The main goal is to detect and recognize periodic behavior in the network traffic within a monitored network infrastructure. Our experiments show that the identified periodic communication and its characteristics can be used to classify the traffic and recognize the communicating application/service/device. This capability is essential in network security, especially for detection of the malicious traffic, e.g., a machine infected by some malware that communicates with Command and Control (C&C) servers. The main advantage of our approach is its applicability even on encrypted traffic since the periodic behavior of the application is usually observable despite the communication was encrypted using standard mechanisms.

II. Related Works

A. Periodicity Detection

Scientists focus on detection and analysis of periodic patterns in time series for many years, and there are innumerable published papers. The reason of such long-lasting research is that each domain or application usually requires some new specifics of time series analysis methods; and there is unfortunately no universal method.

A typically used method to detect periodicity is a periodogram defined by Arthur Schuster in [1]. The periodogram creates a spectral density of a signal and it is used to identify important frequencies of a time series. There are many versions derived from the original periodogram, e.g., Bartlett's procedure [2], Welch method [3], Laplace periodogram [4] and Lomb-Scargle periodogram [5].

Another typical method of detecting periodicity is the autocorrelation function, which is based on checking a correlation of a time series with a lagged version of itself. The origin of correlation was explained by Pearson in the article [6].

There are also methods of detecting periodic behavior that are rather engineering than mathematical, for example, an apriori [7]. This technique of frequent pattern mining use two steps called "join" and "prune" to reduce a searching space and these steps are iterated in loops. The apriori was also improved into periodicity mining [8].

B. Use of Periodic Behavior in Network Traffic

Detection of periodicity in network communication is useful in many areas, for example, anomaly detection [9], [10], traffic filtering [11] or creating model of traffic [12].

Some papers also use identified periodic behavior for a classification. Paper [13] aims to detect periodicity in HTTP traffic filtered to GET and POST methods. A simple decision tree assigns five tags (classes) based on type of the observed periodicity. Based on these tags, the authors decide if an HTTP traffic is generated by a botnet or not. The paper [14] purposes a novel detection model named detection by mining regional periodicity (DMRP), which is used to detection of P2P botnets.

978-3-903176-51-5

Paper [15] shows that periodic traffic analysis is effective for detecting P2P, gaming, cloud, scanning, and botnet traffic flows. For detection periodicity, they are using SQL-based implementation of the periodicity detection method proposed by Hubballi and Goyal [16]. Also, paper [17] uses periodicity as one of the three features to identify webmail traffic.

C. Difference from Related Work

Contrary to the listed related works, our paper targets on the classification of any process that generates communication. Significant groups of classification classes are *social networks*, *remote storage*, *webmails*, *antiviruses*, *operating systems*, *network protocols*, and *network services*. Also, our paper uses mathematical tools and statistics to detect periodic behaviors in network traffic, while the related works uses mainly engineering methods.

Our methodology was inspired by the works such as [14]; contrary, we use periodogram for confirmation of periodicity candidate. Also the idea of identifying candidates on periodicity using the autocorrelation function and periodogram (hybrid method) was shown, e.g., in [18] as an Autoperiod method, and also in [19] as a CDF-Autoperiod method; contrary, we use Lomb-Scargle periodogram instead of the traditional periodogram and we apply it on network traffic. Therefore, we incorporated such workflow into the proposed network classification method based on identified periodic behavior.

III. Our Approach

This section describes several terms to explain our approach to detect periodic communication in network traffic and use it for classification using machine learning.

A. Network Dependencies

At first, we define a *network dependency* as a long-term relationship (e.g., longer than one traditional IP flow) between two IP addresses (devices), where one of them provides some service and the second one communicates with it. Thus, a network dependency is represented by a sequence of communications. We estimate the used service using ports of the transport protocol of the TCP/IP model, and primarily we assume well-known and registered ports identify the services. Example of a network dependency is *192.168.1.1(53)–192.168.1.110* where IP *192.168.1.1* provides a DNS service IP *192.168.1.110* communicates with it.

For unregistered ports, we define the network dependency in multiple ways based on observed communication between two IP addresses. When two ports appear multiple times, we define network dependency as a relationship between two IP addresses with those two ports, e.g., *78.128.191.25(5965-8884)–3.209.182.156*. If only one of the ports appears multiple times and the second one is changing with every flow record then we create a network dependency based on the repeated port, e.g., *138.232.236.27(X-5228)–142.250.27.188*.

Finally, if there is no stable port used for communication between two IP addresses, we define the network dependency only by the IP addresses, e.g., *192.168.1.110()–192.168.1.111*.

B. Time Series from Network Traffic

Network traffic represented by IP flows (IPFIX) from a monitoring system is split among network dependencies and time series are created for each of them. We define a time series for a network dependency as a sequence of flow records f_t, where $t \in T$ is the time of the flow record, and T is a time interval during which we measured on the network. In our experiments, we use number of packets and bytes, which are available in every monitoring system. We create a multivariate time series where each selected value from the flow record is a variable of the multivariate time series.

Time information of the flow record t (timestamp of the first packet) is unevenly spaced, thus the time series are called irregularly spaced. Such time series contains gaps of different length. The example of an unevenly spaced time series is showed in Fig. 1.

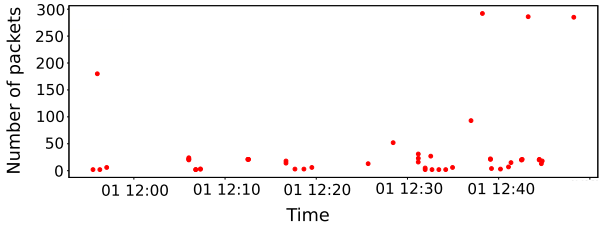

Fig. 1. Time series of network dependency 142.251.36.74(443)-10.0.0.10

Based on the time series progress, we define *periodic network communication* as the repeated transfer of packets that have the same purpose.

C. Periodic Behavior in Time Series from Network Traffic

We identify two types of periodic behavior in the time series from network traffic. The first is a periodic repeating of the same volume of data in some stable time intervals. The time interval can slightly change due to the delay of the source or network components, so we permit small variance in time. An example of this behavior is the time series of the *Mirai* malware communication shown in the Fig. 2.

The second type of behavior has data points of time series distributed in time similarly as the first type, however, the volume of data might vary (e.g., based on the size of the messages' content). This periodic behavior is generated by a process that communicates periodically, for example, due to polling for new state. An example of this behavior is shown in the Fig. 3 where is the time series of the social network *Facebook*.

D. Detection of periodic behavior in Time Series from Network Traffic

To detect both types of periodic behaviors in network traffic, we use mathematical tools to detect periodicity and periodic behavior in the time series. However, the analysis of unevenly sampled time series must compute not only the values of the data points but also their time, and unfortunately, most time series analysis methods specialize in the evenly sampled time

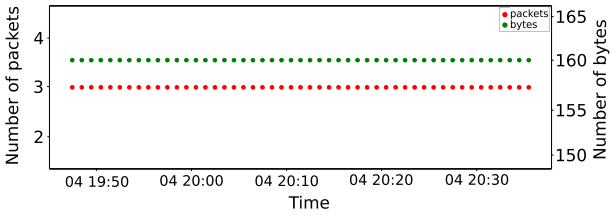

Fig. 2. Periodic behavior of a time series of the malware Mirai

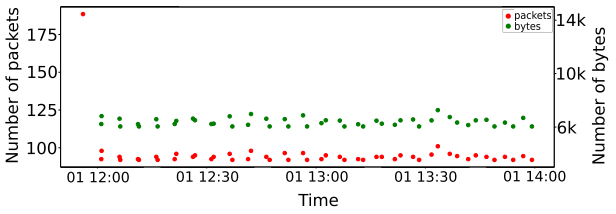

Fig. 3. Periodic behavior of a time series of the social network Facebook

series that appear in most areas. A promising method is a *Lomb-Scargle periodogram* that was defined by Lomb in [20] and Scargle in [21]. It can insert different sine signals into an unevenly sampled time series of periodic behavior and derive frequency and intensity for each, thus creating a periodogram. For the example time series shown in Fig. 4, the Lomb-Scargle periodogram is in Fig. 5.

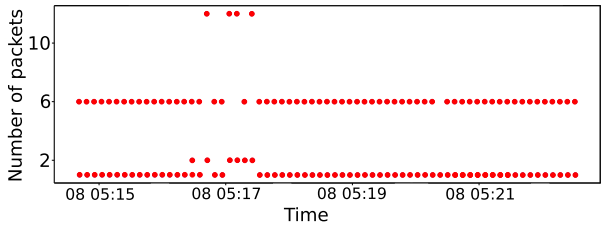

Fig. 4. Time series of network dependency ff02::1:2(547)-fe80::1

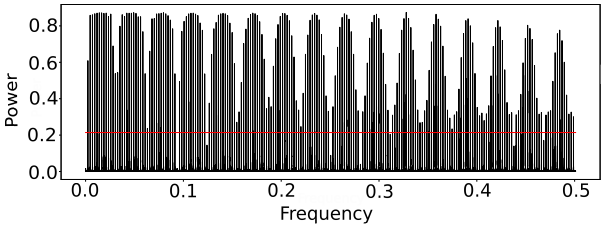

Fig. 5. Lomb-Scargle periodogram of the time series from network traffic from the Fig. 4

The time series is periodic with period p, if there is a statistically significant peak at frequency $f = \frac{1}{p}$. It is therefore necessary to use a statistical significance test on the periodogram. The *Scargle's Cumulative Distribution Function (SCDF)* [22] is suitable for the Lomb-Scargle periodogram, which can be used to determine whether there is any statistically significant peak in the periodogram using a simple formula and also to find out whether a particular peak is statistically significant. In Fig. 5 the red line represent the *SCDF* test of significance.

In order to confirm the periodicity of candidate p, we must first know what candidates make sense to test. For this purpose, we use an autocorrelation function that does not work with unevenly sampled time series, so when we use it we neglect time. An example of applying the autocorrelation function to the variable number of packets of the time series from network traffic shown in the Fig. 4 can be seen in the Fig. 6.

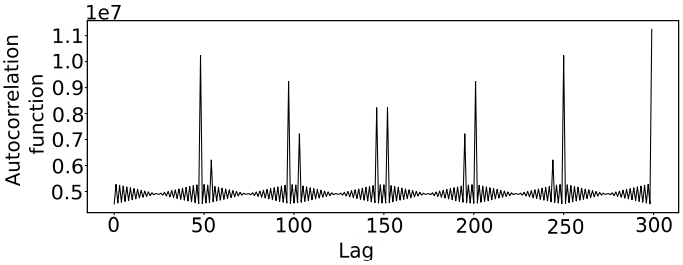

Fig. 6. Autocorrelation function of the time series variable number of packets from the Fig. 4

Using the autocorrelation functions computed for each variable of the time series (i.e., number of packets, number of bytes, DiffTimes — a time difference between intermediate data points of time series), we create histograms of distances between the peaks. Such histogram per each variable is shown in Fig. 7.

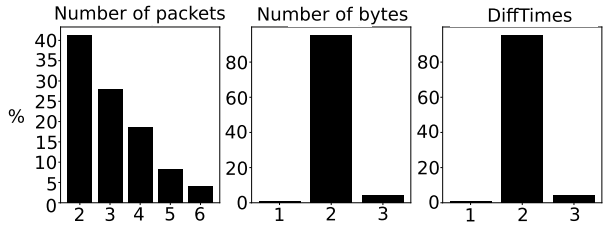

Fig. 7. Histograms of distances between peaks of the autocorrelation functions applied on variables of the time series

The result is a set of candidates for periodicity that significantly appear in all histograms, which can be verified subsequently using the Lomb-Scargle periodogram and the SCDF test. In this particular example, period 2 is chosen as the candidate and a statistically significant peak is then found in the vicinity of the frequency 0.5 and so the period is confirmed.

In general, network dependencies may not always split traffic into ideal time series, thus, there may be multiple periodic behaviors in a single time series. Therefore, we need to check whether there is a possibility that the time series does not contain other periodic behavior. The diagram of the whole method is shown in the Fig. 8.

E. Attributes of periodic behavior

A natural description of periodic behavior from Fig. 2 is a *number of flow records* that periodically repeat, a *time period*, a *number of packets*, and a *number of bytes in the flow*.

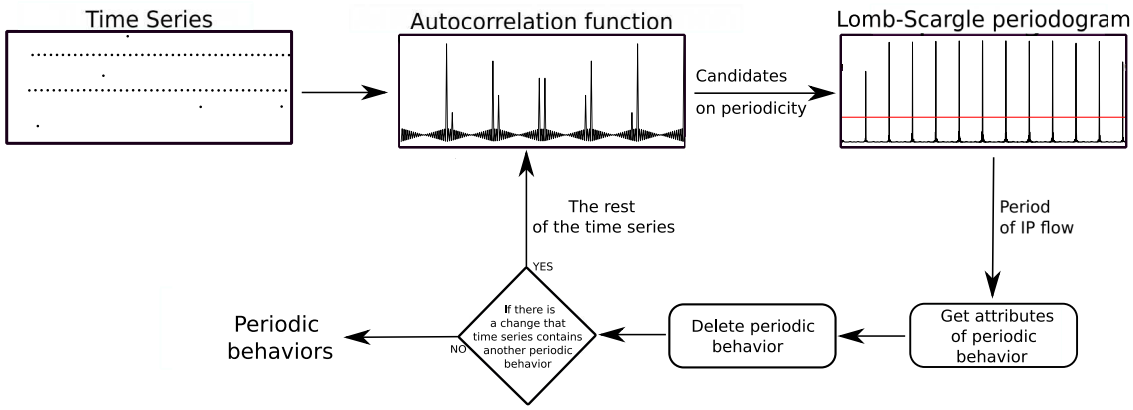

Fig. 8. Diagram of periodicity detection method

However, these attributes cannot perfectly describe a different periodic behavior from Fig. 3 because of its higher variability. Therefore, we use *boundaries of the interval* of the data points for this type of behavior, whereas the boundaries embody the most of data points for each time series variable.

IV. Dataset

We have created a dataset containing 26 thousands samples divided into 61 classification classes. As a traffic source for the dataset creation we used multiple publicly available datasets: [23], [24], [25] and [26]. Additionally, we used traffic of CESNET2[1], traffic of Network monitoring laboratory at FIT Czech Technical University in Prague, several home networks, and communication of Android mobile devices.

The following traffic categories list the examples of the classes for classification (listed in brackets):

- social networks (Facebook, MS teams, Slack, ...)
- remote storage (Google Drive, OneDrive, Github, ...)
- updates of operating systems
- antivirus programs (Eset, Avast, Kaspersky, ...)
- game clients (Steam, Epic Games, Uplay, ...)
- network services and protocols (Keep-alive, HTTP2 ping, DNS, ...)
- email browser viewers and clients (Gmail, Outlook, ...)
- multimedia streaming (youtube, itunes, spotify)
- web browsing (firefox, opera)

V. Machine Learning Models

Traffic characteristics from IP flow records with the results of our periodicity analysis allow for training new classification models based on machine learning to recognize classes of traffic mentioned in Sec. IV.

More specifically, we evaluated the list of algorithms (shown in Tab. I) using the *time period*, *number of packets and bytes*, *boundaries of the interval of packets and bytes* (attributes of periodicity computed according to Sec. III) as features for the machine learning. Even though the time period is not the most important feature (according to the computed feature importance metric), it is still helpful for classifiers and thus included in the feature vector. For the first type of periodic behavior, i.e., communication is regular without significant variance, we set the boundaries of the interval of packets and bytes to 0. The dataset was split into training and testing parts in a ratio of 70 to 30. The best performing classification algorithm was XGBoost with F1-score 90%. Using *K-fold validation*, we achieved 86% accuracy of the selected model.

TABLE I
Results of classification algorithms

Model	Accuracy	Precision	Recall	F1-score
Naive Bayes	8	25	8	4
Logistic Regression	42	19	42	26
kNN	62	62	62	61
Extra Tree	65	66	66	66
Decision Tree	77	77	77	77
Random Forest	89	90	89	88
XGBoost	91	90	90	90

Results of our experiments showed that network traffic of the second type of periodicity contains, e.g., social networks communication. This type is harder to recognize, as it is shown in the confusion matrix in Fig. 9. Contrary, system/application level traffic such as keepalive, polling etc. fits to the first type of periodicity. In this case, the classifier is able to recognize each class with higher accuracy. Additionally, the more often the application communicates, the higher accuracy. The results are shown in the confusion matrix in Fig. 10.

VI. Conclusion

Encrypted network traffic rises new challenges in network monitoring and network security domains. Luckily, there are some behavioral patterns and characteristics that reveal information about the communication and can be exploited as a "side-channel" without any need to decrypt the content.

This paper elaborated on detection of periodic behavior of network traffic and applications. Even though it is a unique area, we have found suitable mathematical tools from astrophysics to discover periodicity in the network traffic. Furthermore, we studied the features of periodic traffic and

[1]the Czech national research and education network

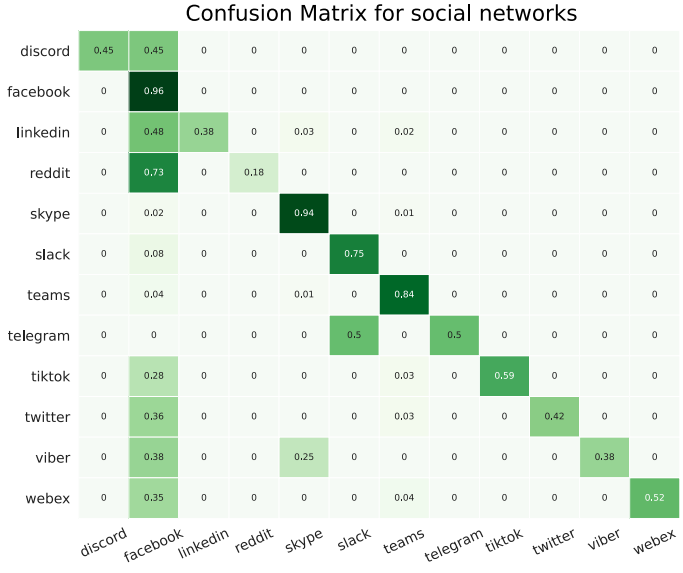

Fig. 9. The confusion matrix for social networks

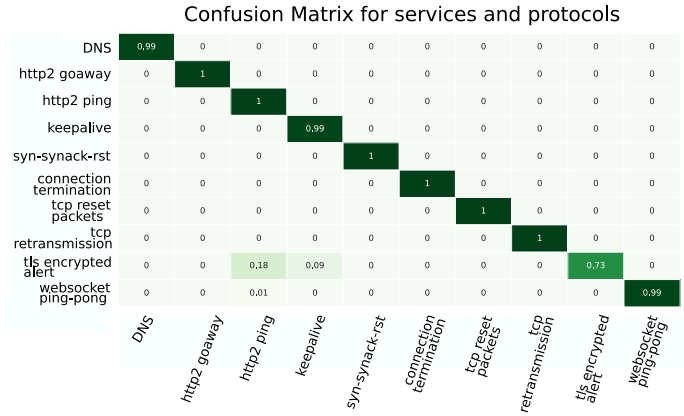

Fig. 10. The confusion matrix for services and protocols

identified two types of it in practice. Finally, we were able to train a classifier based on machine learning that exploits traffic statistics and reveals information about periodic behavior to recognize particular application that originated it. In total, we were able to classify 61 types of traffic with F1-score 90%.

As a future work, we will focus on the applicability of the proposed classifier in high-speed networks. That means the performance will be improved to retrieve the results faster from the classifier. Also, we believe further study of the attributes of periodicity can increase accuracy. Finally, handling of unknown classes would be useful for practical deployment.

Acknowledgment

This research was funded by the Ministry of Interior of the Czech Republic, grant No. VJ02010024: Flow-Based Encrypted Traffic Analysis and also by the Grant Agency of the CTU in Prague, grant No. SGS20/210/OHK3/3T/18 funded by the MEYS of the Czech Republic.

References

[1] Arthur Schuster, Henry Ludwell Moore, and A. E. Douglass. *Periodogram analysis*. 1898.
[2] M. S. Bartlett. Periodogram analysis and continuous spectra, Jun 1950.
[3] Kurt Barbe, Rik Pintelon, and Johan Schoukens. Welch method revisited: Nonparametric power spectrum estimation via circular overlap. *IEEE Transactions on Signal Processing*, 58(2):553–565, 2010.
[4] Ta-Hsin Li. Laplace periodogram for time series analysis. *Journal of the American Statistical Association*, 103(482):757–768, 2008.
[5] Jacob T. VanderPlas. Understanding the lomb–scargle periodogram. *The Astrophysical Journal Supplement Series*, 236(1):16, may 2018.
[6] Karl Pearson. Notes on the history of correlation. *Biometrika*, 13(1):25–45, 1920.
[7] Rakesh Agrawal, Ramakrishnan Srikant, et al. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, volume 1215, pages 487–499. Santiago, Chile, 1994.
[8] Jiawei Han, Guozhu Dong, and Yiwen Yin. Efficient mining of partial periodic patterns in time series database. In *Proceedings 15th International Conference on Data Engineering*. IEEE, 1999.
[9] Genevieve Bartlett et al. Using Low-Rate Flow Periodicities for Anomaly Detection: Extended. Technical Report ISI-TR-2009-661, August 2009.
[10] Rafael Ramos Regis Barbosa, Ramin Sadre, and Aiko Pras. Towards periodicity based anomaly detection in scada networks. In *Proceedings of 2012 IEEE 17th International Conference on Emerging Technologies and Factory Automation (ETFA 2012)*, pages 1–4, 2012.
[11] Francesco Sanna Passino and Nicholas A Heard. Classification of periodic arrivals in event time data for filtering computer network traffic. *Statistics and Computing*, 30(5):1241–1254, 2020.
[12] Rafael Ramos Regis Barbosa, Ramin Sadre, and Aiko Pras. Exploiting traffic periodicity in industrial control networks. *International Journal of Critical Infrastructure Protection, 13*, 2016.
[13] Meisam Eslahi, M. S. Rohmad, Hamid Nilsaz, Maryam Var Naseri, N.M. Tahir, and H. Hashim. Periodicity classification of http traffic to detect http botnets. In *2015 IEEE Symposium on Computer Applications and Industrial Electronics (ISCAIE)*, pages 119–123, 2015.
[14] Yong Qiao et al. Detecting P2P bots by mining the regional periodicity. *Journal of Zhejiang University SCIENCE C*, 14(9):682–700, Sep 2013.
[15] Mackenzie Haffey et al. Modeling, analysis, and characterization of periodic traffic on a campus edge network. In *2018 IEEE 26th International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS)*, 2018.
[16] Neminath Hubballi and Deepanshu Goyal. Flowsummary: Summarizing network flows for communication periodicity detection. In *Pattern Recognition and Machine Intelligence*, 2013.
[17] Dominik Schatzmann et al. Digging into HTTPS: flow-based classification of webmail traffic. 2010.
[18] Michalis Vlachos, Philip Yu, and Vittorio Castelli. On periodicity detection and structural periodic similarity. 04 2005.
[19] Tom Puech et al. A fully automated periodicity detection in time series. In *Advanced Analytics and Learning on Temporal Data*, 2020.
[20] N.R. Lomb. Least-squares frequency analysis of unequally spaced data. 1976.
[21] Jeffrey Scargle. Studies in astronomical time series analysis. ii - statistical aspects of spectral analysis of unevenly spaced data. *The Astrophysical Journal*, 263, December 1982.
[22] Fabio Frescura, Chris Engelbrecht, and B. Frank. Significance tests for periodogram peaks. 2007.
[23] Romain Fontugne et al. Mawilab: Combining diverse anomaly detectors for automated anomaly labeling and performance benchmarking. In *ACM CoNEXT '10*, December 2010.
[24] Manmeet; Singh Singh, Maninder; Kaur, and Sanmeet. *"10 Days DNS Network Traffic from April-May, 2016", Mendeley Data, V2*. doi: 10.17632/zh3wnddzxy.2.
[25] G. Creech and J. Hu. *ADFA IDS Dataset, University of Arizona Artificial Intelligence Lab, AZSecure-data*. Director Hsinchun Chen, November 2016. http://www.azsecure-data.org/.
[26] Student Union for Electrical Engineering (Fachbereichsvertretung Elektrotechnik) at Ulm University and Philipp Hinz. *2017 SUEE data set*. https://github.com/vs-uulm/2017-SUEE-data-set, Accessed: 2022-06-24.