# Feature description of time series features of datasets for the article “Network Traffic Classification based on Single Flow Time Series Analysis”

1st Josef Koumar
*Czech Technical University in Prague*
Prague, Czech republic
koumajos@fit.cvut.cz

2nd Karel Hynek
*Czech Technical University in Prague*
Prague, Czech republic
hynekkar@fit.cvut.cz,

3rd Tomas Čejka
*CESNET, a.l.e.*
Prague, Czech republic
cejkat@cesnet.cz

Tables I, II, III, IV, and V present features with textual descriptions and mathematical equations for the ability to recompute the time series features. The mathematical equations have the following notation for one IP flow, i.e., one network communication between a server and client:

$\{X_n\}$ is a Single flow time series (SFTS), i.e., sequence of $n$ payload lengths of network packets $\{x_0, \ldots, x_{n-1}\}$ within one IP flow, and $x_i$ is the $i$-th value.

$\{t_n\}$ is sequence of times of SFTS, $\{X_n\}$, i.e. $i$-th packet is transferred in the $t_i \in \{t_n\}$.

$\{x'\}$ is a sequence of payload lengths of each network packet sorted by value in ascending order, and $x'_i$ is $i$-th value.

$\{st_n\}$ is a sequence of Scaled times computed by the equation: $st_i = t_i - t_0, i \in \{0, \ldots, n-1\}$.

$\{st'\}$ is a sequence of Scaled times sorted by value in ascending order, and $st'_i$ is $i$-th value.

$\{dt_{n-1}\}$ is a sequence of Time differences, i.e., spaced between observations, computed by the equation: $dt_i = t_{i+1} - t_i, i \in \{0, \ldots, n-2\}$.

$\{dt'\}$ is a sequence of Time differences sorted by value in ascending order, and $dt'_i$ is $i$-th value.

$\{d\}$ is a sequence of payload lengths that occur in time series sorted by value in descending order, and $d_i$ is $i$-th value. Only unique payload lengths are included.

$\{c\}$ is a sequence of the number of occurrences of payload length sorted in descending order, and $c_i$ is $i$-th value and it is a number of occurrences of payload length $d_i$.

$\{y_m\}$ is aggregated SFTS on 1-second intervals, and $y_i$ is $i$-th value of the time series of $m$ values.

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. SGS23/207/OHK3/3T/18 funded by the MEYS of the Czech Republic.

$\{z_k\}$ is sequence of non-zero values of $\{y\}$, and $z_i$ is $i$-th value of time series of $k$ values.

$\{P_{LS}\}$ is the power spectrum of the Lomb-Scargle (LS) periodogram [1]–[3]. The $P_{LS}(f_j)$ is power on frequency $f_j \in \{f\}$. The generalized form of the LS periodogram for an unevenly spaced time series $\{x_n\}$ with times $\{t_n\}$ is shown in equation 1.

$$P_{LS}(f_j) = \frac{1}{2}\frac{\left(\sum_i x_i cos\left(2\pi f_j [t_i - \tau]\right)\right)^2}{\sum_i cos^2\left(2\pi f_j [t_i - \tau]\right)} + \frac{1}{2}\frac{\left(\sum_i x_i sin\left(2\pi f_j [t_i - \tau]\right)\right)^2}{\sum_i sin^2\left(2\pi f_j [t_i - \tau]\right)} \quad (1)$$

where $\tau$ is specified for each frequency $f_j$ to ensure time-shift invariance:

$$\tau = \frac{1}{4\pi f_j} tan^{-1}\left(\frac{\sum_i sin\left(4\pi f_j t_i\right)}{\sum_i cos\left(4\pi f_j t_i\right)}\right) \quad (2)$$

$\{f\}$ is sequence of frequencies for which there is a power in LS periodogram $\{P_{LS}\}$, and $f_j \in \{f\}$ is $j$-th frequency.

$\{\hat{f}\}$ is a sequence of frequencies in reverse order for which there is a power in LS periodogram $\{P_{LS}\}$, and $\hat{f}_j \in \{\hat{f}\}$.

$N$ is the number of frequencies of the Lomb-Scargle periodogram.

## References

[1] Nicholas R Lomb. Least-squares frequency analysis of unequally spaced data. *Astrophysics and space science*, 39:447–462, 1976.

[2] Jeffrey Scargle. Studies in astronomical time series analysis. II - Statistical aspects of spectral analysis of unevenly spaced data. *The Astrophysical Journal*, 263, December 1982.

[3] Jacob T. VanderPlas. Understanding the lomb–scargle Periodogram. *The Astrophysical Journal Supplement Series*, 236(1):16, may 2018.

[4] A. A. Anis and E. H. Lloyd. The Expected Value of the Adjusted Rescaled Hurst Range of Independent Normal Summands. *Biometrika*, 63(1), 1976.

[5] Denis Kwiatkowski et al. Testing the Null Hypothesis of Stationarity Against the Alternative of a Unit Root: How Sure Are We That Economic Time Series Have a Unit Root? *Journal of Econometrics*, 54(1), 1992.

[6] Steven J Miller. *Benford's Law*. Princeton University Press, 2015.

[7] Thomas L. Szabo. 5 - Transducers. In Thomas L. Szabo, editor, *Diagnostic Ultrasound Imaging*, Biomedical Engineering, pages 97–135. Academic Press, Burlington, 2004.

[8] Eric D. Scheirer and Malcolm Slaney. Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator. In *ICASSP 1997*, pages 1331–1334. IEEE Computer Society, 1997.

[9] Boualem Boashash. *Time-Frequency Signal Analysis and Processing: A Comprehensive Reference*. Academic press, 2015.

[10] Valeriu Vrabie, Pierre Granjon, and Christine Serviere. Spectral Kurtosis: From Definition to Application. In *6th IEEE international workshop on Nonlinear Signal and Image Processing (NSIP 2003)*, page xx, 2003.

[11] FAM Frescura, CA Engelbrecht, and BS Frank. Significance Tests for Periodogram Peaks. *arXiv preprint arXiv:0706.2225*, 2007.

[12] Md Gulzar Hussain et al. Classification of bangla alphabets phoneme based on audio features using mlpc & svm. In *ACMI 2021*. IEEE, 2021.

[13] Geoffroy Peeters. A Large Set of Audio Features for Sound Description (Similarity and Classification) in the CUIDADO Project. *CUIDADO Ist Project Report*, 54(0):1–25, 2004.

[14] Alexander Lerch. *An Introduction to Audio Content Analysis: Applications in Signal Processing and Music Informatics*. Wiley-IEEE Press, 2012.

TABLE I
SUMMARY DETAILED DESCRIPTION OF STATISTICAL-BASED FEATURES OF THE NETTISA FLOW

| Feature | Mathematical equation | Description |
|---|---|---|
| **Mean** | $\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$ | The average value of data points |
| **Median** | $\widetilde{x} = x'_{\frac{n+1}{2}}$ | The middle value of sorted data points |
| **Standard deviation** | $\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2}$ | The measure of the variation of data from the mean. |
| **Variance** | $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$ | The measure of spread of data from its mean. |
| **Percent above mean** | $= \frac{\alpha}{n}$, where $\alpha$ is the number of values greater than $\mu$ | Percent of data points with value greater than the mean |
| **Percent below mean** | $= \frac{\beta}{n}$, where $\beta$ is the number of values lower than $\mu$ | Percent of data points with value smaller the mean |
| **Burtiness** | $b_{x_n} = \frac{\sigma-\mu}{\sigma+\mu}$ | The degree of peakedness in the central part of the distribution. |
| **First quartile** | $Q1_{\{x_n\}} = x'_{\frac{n+1}{4}}$ | The value marking off the highest 25% of values. |
| **Third quartile** | $Q3_{\{x_n\}} = x'_{\frac{3(n+1)}{4}}$ | The value marking off the highest 75% of values. |
| **Min** | $\min(x_1, x_2, \ldots, x_n)$ | Minimum value in the time series. |
| **Max** | $\max(x_1, x_2, \ldots, x_n)$ | Maximum value in the time series. |
| **Min minus max** | $= \lvert\min - \max\rvert$ | Difference between minimum and maximum in the time series. |
| **Mode** | $M_o = \text{argmax}(f(x_1), \ldots, f(x_n))$ | Most common value in the time series. |
| **Average dispersion** | $ad = \frac{1}{n}\sum_{i=1}^{n}\lvert x_i-\mu\rvert$ | The average absolute difference between each data point and the mean value of the time series. |
| **Percent deviation** | $pd = \frac{ad}{\mu}$ | The dispersion of the average absolute difference to the mean value. |
| **Root mean square** | $\text{rms} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}$ | The measure of the magnitude of a time series values. |
| **Entropy** | $H(\{x_n\}) = -\sum_{i=1}^{n} p_i \log_2 p_i$ | The measure of the amount of uncertainty or randomness in the time series. |
| **Scaled entropy** | $H_s(\{x_n\}) = \frac{H(\{x_n\})}{-\log_2 \frac{1}{n}}$ | It normalizes the entropy value by dividing it by the logarithm of the time series length, and it is often used to compare the entropy values of time series with different lengths. |
| **Kurtosis** | $\text{kurt} = \frac{1}{n\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^4$ | The measure describing the extent to which the tails of distribution differ from the tails of a normal distribution. |
| **Coefficient of variation** | $\text{cv} = \frac{\sigma}{\mu}$ | The dimensionless quantity that compares the dispersion of a time series to its mean value and is often used to compare the variability of different time series that have different units of measurement. |
| **Galton skewness** | $G_s = \frac{Q_1+Q_3-2\mu}{Q_3-Q_1}$ | The measure of the asymmetry of a probability distribution that is based on the difference between the arithmetic mean and the median of the time series. It is less sensitive to outliers than other measures of skewness. It is often used in financial analysis and risk management. |
| **Pearson SK$_1$ skewness** | $\text{sk}_1 = \frac{\mu-M_o}{\sigma}$ | The measure of the skewness based on the standardized third central moment of the time series, and it is often used in statistical analysis and modelling. It is more sensitive to outliers. |
| **Pearson SK$_2$ skewness** | $\text{sk}_2 = \frac{3\mu-\widetilde{x}}{\sigma}$ | It is a commonly used measure of skewness due to its ability to detect both positive and negative skewness. |
| **Fisher $\mu_3$ skewness** | $\mu_3 = E\left[\left(\frac{\{x_n\}-\mu}{\sigma}\right)^3\right]$ | It is designed to be unbiased, meaning that it estimates the true skewness of a population based on a sample of data without over- or underestimating it. |
| **Fisher-Pearson $g_1$ skewness** | $g_1 = \frac{1}{\sigma^3}\sum_{i=1}^{n}\frac{(x_i-\mu)^3}{n}$ | It is designed to have a value of zero for symmetrical distributions, which makes it a more suitable measure for comparing skewness across different types of distributions. |
| **Fisher-Pearson $G_1$ skewness** | $G_1 = \frac{\sqrt{n(n-1)}}{n-2} g_1$ | It is designed to have a value of zero for symmetrical distributions, making it a more appropriate measure for comparing skewness across different types of distributions. |

TABLE II
SUMMARY DETAILED DESCRIPTION OF TIME-BASED FEATURES OF THE NETTISA FLOW

| Feature | Mathematical equation | Description |
|---|---|---|
| **Mean of scaled times** | $\mu_{st} = \frac{1}{n}\sum_{i=1}^{n} st_i$ | The average value of data points |
| **Median of scaled times** | $\widetilde{st} = st'_{\frac{n+1}{2}}$ | The middle value of sorted data points |
| **First quartile of scaled times** | $Q1_{\{st_n\}} = st'_{\frac{n+1}{4}}$ | The value marking off the highest 25% of values. |
| **Third quartile of scaled times** | $Q3_{\{st_n\}} = st'_{\frac{3(n+1)}{4}}$ | The value marking off the highest 75% of values. |
| **Mean of time differences** | $\mu_{dt} = \frac{1}{n}\sum_{i=1}^{n} dt_i$ | The average value of data points |
| **Median of time differences** | $\widetilde{dt} = dt'_{\frac{n+1}{2}}$ | The middle value of sorted data points |
| **First quartile of time differences** | $Q1_{\{dt_n\}} = dt'_{\frac{n+1}{4}}$ | The value marking off the highest 25% of values. |
| **Third quartile of time differences** | $Q3_{\{dt_n\}} = dt'_{\frac{3(n+1)}{4}}$ | The value marking off the highest 75% of values. |
| **Duration** | $D = t_{n-1} - t_0$ | The time duration of the IP flow. |

TABLE III
SUMMARY DETAILED DESCRIPTION OF DISTRIBUTION-BASED FEATURES OF THE NETTISA FLOW

| Feature | Mathematical equation | Description |
|---|---|---|
| **Hurst exponent** | $E\left[\frac{R(n)}{S(n)}\right] = Cn^H$ where $R(n)$ is the first $n$ cumulative deviations from the mean, $S(n)$ is the sum of the first $n$ standard deviations, $E$ is the expected value, and $C$ is a constant. | The Hurst exponent $H$ can detect a time series's tendency to regress to the mean or cluster towards the center strongly. If $H \in \langle 0; 0.5)$, it indicates a long-term switching between high and low values in adjacent pairs. It is also stated that the time series is anti-persistent. If $H \sim 0.5$, then this indicates a random (uncorrelated) time series. Furthermore, if $H \in (0.5; 1\rangle$ indicates a long-term positive autocorrelation in the time series. It is also said that the time series is persistent. [4] |
| **Stationarity** | Adfuler test of stationarity | Properties of a stationary time series do not depend on the observation time. So time series with a trend or with seasonality is not stationary. Nevertheless, the time series with periodic (or cyclic) behavior without trend or seasonality is stationary. [5] |
| **Benford's law** | $P_{\text{BENFORD}} = 1 - \frac{1}{2}\sum_{i=1}^{9}\left(\log_{10}\left(1+\frac{1}{d_i}\right) - \frac{c_i}{n}\right)$ | Describes a probability that the occurrence counts of the first 9 most frequent data points of the time series conform to Benford's law [6]. |
| **Normal distribution** | Lilliefors test of normality | Verify if the aggregated SFTS to 1-second intervals is distributed by the normal distribution. That means deciding if most of the communication of the flow is suited in the middle of the flow. |
| **Count distribution** | $\text{cdist} = \frac{\frac{1}{m}\sum_{i=1}^{m}\left\|\mu_{\{y_m\}} - y_i\right\|}{\frac{1}{2}(max(\{y_m\}) - min(\{y_m\}))}$ | Describes the distribution of the number of packets over time. The lower $cdist$ is, the better the packet counts are distributed across the time series. The disadvantage is that the zero intervals can dominate and artificially reduce the value when there are many zero intervals. |
| **Count non-zero distribution** | $\text{cndist} = \frac{\frac{1}{k}\sum_{i=1}^{k}\left\|\mu_{\{z_k\}} - z_i\right\|}{\frac{1}{2}(max(\{z_k\}) - min(\{z_k\}))}$ | This feature is similar to feature *Count distribution* but filters the data points with zero value out of aggregated time series. |
| **Time distribution** | $\text{tdist} = \frac{\frac{1}{n-1}\sum_{i=1}^{n-1}\left\|\mu_{\{dt_{n-1}\}} - dt_i\right\|}{\frac{1}{2}\left(max\left(\{dt_{n-1}\}\right) - min\left(\{dt_{n-1}\}\right)\right)}$ | Describes the distribution of time differences between individual packets. The lower the $tdist$, the better the time differences are spread over time. The weakness is, for example, if there are only two values in the time series $\{dt_{n-1}\}$ with the same count, then the result is always 1 because the mean, $\mu_{\{dt_{n-1}\}}$, will always be exactly between them, and so denominator and numerator will have the same value. |

TABLE IV
SUMMARY DETAILED DESCRIPTION OF FREQUENCY-BASED FEATURES OF THE NETTISA FLOW

| Feature | Mathematical equation | Description |
|---|---|---|
| **Min power** | $min\left(\{P_{LS}\}\right)$ | The minimum power of the LS periodogram. |
| **Max power** | $max\left(\{P_{LS}\}\right)$ | The maximum power of the LS periodogram. |
| **Frequency of min power** | $f_j \mid P_{LS}(f_j) == min\left(\{P_{LS}\}\right)$ | The frequency of the *Min power*. |
| **Frequency of max power** | $f_j \mid P_{LS}(f_j) == max\left(\{P_{LS}\}\right)$ | The frequency of the *Max power*. |
| **Power mean** | $\mu_{LS} = \frac{1}{N}\sum_{f_j \in \{f\}} P_{LS}(f)$ | The average power of the LS periodogram |
| **Power mode** | $M_{P_{LS}} = argmax(\{P_{LS}\})$ | Most common power in the LS periodogram. |
| **Power standard deviation** | $\sigma_{LS} = \sqrt{\frac{1}{N}\sum_{f_j \in \{f\}} \left(P_{LS}(f_j) - \mu_{LS}\right)^2}$ | The measure of the variation of powers from the power mean. |
| **Spectral bandwidth** | $S_b = \sum_{f_j \in \{f\}} P_{LS}(f_j)\left(f_j - S_c\right)^{\frac{1}{p}}$ | Computes the order-$p$ spectral bandwidth that aims to describe the difference between upper and lower frequencies at which spectral energy is half its maximum value [7]. |
| **Spectral centroid** | $S_c = \frac{\sum_{f_j \in \{f\}} f_j P_{LS}(f_j)}{\sum_{f_j \in \{f\}} P_{LS}}$ | Indicates at which frequency the energy of a spectrum is centred upon [8]. |
| **Spectral energy** | $S_e = \sum_{f_j \in \{f\}} P_{LS}(f_j)$ | Represents the total energy present at all frequencies in LS periodogram |
| **Spectral entropy** | Computation is same as for classic entropy but each $p_i$ is a probability of some power on LS periodogram. | The degree of randomness or disorder in the LS periodogram. |
| **Spectral flatness** | $S_{f_j} = \frac{\sqrt[N]{\Pi_{f_j \in \{f\}} P_{LS}(f_j)}}{\frac{1}{N}\sum_{f_j \in \{f\}} P_{LS}(f_j)}$ | Estimate the uniformity of signal energy distribution in the frequency domain [9]. (sometimes called a spectral crest) |
| **Spectral flux** | $S_F = \left(\sum_{f_j \in \{f\}, \hat{f}_j \in \{\hat{f}\}} \left\lvert P_{LS}(f_j) - P_{LS}(\hat{f}_j)\right\rvert\right)$ | the rate of change of periodogram power with increasing frequency [8] |
| **Spectral kurtosis** | $S_K = \frac{\sum_{f_j \in \{f\}} f_j^4}{\left(\sum_{f_j \in \{f\}} f_j^2\right)^2} - 3$ | Can indicate a nonstationary or non-Gaussian behavior in the power spectrum [10]. |
| **Spectral periodicity** | $SCDF = 1000 - E\left[\frac{-M_{P_{LS}}}{\sigma^2_{P_{LS}}}\right]$ The SFTS contains a periodic signal if it is true $SCDF < t$, where $t$ is a threshold that can be set. From our experiments, we set $t = 0.9995$. Then the feature is set to *True*, otherwise is set to *False*. | The goal of this feature is to decide if in Lomb-Scargle periodogram is a significant peak that indicates the presence of the periodic signal in the SFTS. We use a test by Scargle's cumulative distribution function (SCDF) [11] to decide if the maximum periodogram power is a significant peak. |
| **Spectral rolloff** | $\{\widetilde{f}\} = \{f_j \mid P_{LS}(f_j) > 0.85 * M_{P_{LS}}\}$ $S_r = \widetilde{f}_0$ | Defined as frequency bellow at which 85% of the distribution power is concentrated [12]. |
| **Spectral spread** | $S_{sp} = \sqrt{\frac{\sum_{f_j \in \{f\}}(f_j - S_c)^2 P_{LS}(f_j)}{\sum_{f_j \in \{f\}} P_{LS}(f_j)}}$ | The difference between highest and lowest frequency in power spectrum [13]. |
| **Spectral skewness** | $S_{sk} = \frac{\sum_{f_j \in \{f\}}(f_j - S_c)^3 P_{LS}(f_j)}{S_{sp}^3 \sum_{f_j \in \{f\}} P_{LS}(f_j)}$ | The measure of peakedness or flatness of power spectrum [13]. |
| **Spectral slope** | $S_{sl} = \frac{\sum_{f_j \in \{f\}}(f_j - \mu_f)(f_j - \mu_{P_{LS}})}{\sum_{f_j \in \{f\}}(f_j - \mu_f)^2}$ | The slope of power spectrum trend in given frequency range [14]. |
| **Spectral zero crossing rate** | zcr $= \frac{1}{N-1}\sum_{f_j \in \{f\}, \hat{f}_j \in \{\hat{f}\}} 1_{R<0}(P_{LS}(f_j), P_{LS}(\hat{f}_j))$, where $f_j$ and $\hat{f}_j$ are adjacent frequencies, and $1_{R<0}(P_{LS}(f_j), P_{LS}(\hat{f}_j))$ is 1 when change from negative to positive in frequencies $f_j$ and $\hat{f}_j$ is observed, otherwise it is 0. | Refers to the rate of shift of the sign of a wave, which is the rate of change from negative to positive or the reverse [12]. |

TABLE V

SUMMARY DETAILED DESCRIPTION OF BEHAVIOR-BASED FEATURES OF THE NETTISA FLOW

| Feature | Mathematical equation | Description |
|---|---|---|
| **Significant spaces** | $\mathcal{S} = \{s_i \| s_i > \mu_{\{dt_{n-1}\}} * (1+t)\ \&\ s_i > \sigma_{\{dt_{n-1}\}} * (1+t), s_i \in \{df_{n-1}\}\}$ | The goal of this feature is to verify if in the SFTS are present some spaces, i.e. time differences, that are significantly bigger than the mean. |
| **Switching ratio** | $sr = \frac{s_n}{\frac{1}{2}(n-1)}$, where $s_n$ is the number of switches | Represents a switching ratio between different values of the sequence of observation. |
| **Transients** | Aims to verify if there is at least one transient in the SFTS. The transient in time series is the behavior when a set of data points occurring in a short time window has significantly larger values than the rest of the data points. | |
| **Count of zeros** | $c_0 = \frac{m-k}{m}$ | Represents a percentage representation of zero value data points of aggregated time series, $\{y_m\}$, from the SFTS to 1-second intervals. |
| **Biggest interval** | $max(\{y_m\})$ | Represents the maximum value of data point of aggregated time series. |
| **Directions** | Describe a percentage ratio of packet direction. If they are all in the direction of 1, then the percentages should be 100%, and if they are all in the direction of -1, then the percentages should be 0%. | |
| **Periodicity** | The length and time of periodically occurring packet, if some are present. | |