Key Statistics Calculated

1. Mean (MB/s):

* Definition: The average bandwidth.
* Significance: Indicates the typical performance of the relay over the analyzed period.

1. Standard Deviation (MB/s):

* Definition: A measure of the amount of variation or dispersion in the bandwidth values.
* Significance: Shows how much the bandwidth varies from the mean. A higher standard deviation means more variability.

1. Range (MB/s):

* Definition: The difference between the maximum and minimum bandwidth values.
* Significance: Provides a sense of the spread of the bandwidth values, indicating the extent of fluctuations.

1. Coefficient of Variation:

* Definition: The ratio of the standard deviation to the mean.
* Significance: Indicates the relative variability of the bandwidth. A higher coefficient of variation means more variation relative to the mean.
* I think this is a very valuable reading because it is a relative measurement.

1. Median (MB/s):

* Definition: The middle value of the bandwidth data when ordered from lowest to highest.
* Significance: Provides a measure of central tendency that is less affected by extreme values than the mean.

1. IQR (Interquartile Range) (MB/s):

* Definition: The range between the 25th percentile (Q1) and the 75th percentile (Q3) of the data.
* Significance: Represents the spread of the middle 50% of the data, giving a sense of the typical range of bandwidth values.

1. Skewness:

* Definition: A measure of the asymmetry of the data distribution.
* Significance: Indicates whether the data are skewed to the left (negative skew) or right (positive skew). Symmetrical data has a skewness of zero.

1. Kurtosis:

* Definition: A measure of the "tailedness" of the data distribution.
* Significance: Indicates the presence of outliers. Higher kurtosis means more outliers.

1. Frequency of Outliers:

* Definition: The number of data points that fall outside the typical range (beyond 1.5 times the IQR from Q1 and Q3).
* Significance: Highlights unusual bandwidth values, which could indicate performance issues or anomalies.

1. ACF (Autocorrelation Function):

* Definition: Measures the correlation between the current bandwidth value and its past values at different time lags.
* Significance: Helps identify patterns and periodicities in the data.
* Example: If you have bandwidth data recorded every hour, ACF will show how similar the bandwidth is to its values from 1 hour ago, 2 hours ago, etc.
* Use Case: Useful for detecting regular fluctuations or cycles in bandwidth, which could indicate recurring performance issues.

1. PACF (Partial Autocorrelation Function):

* Definition: Measures the correlation between the current bandwidth value and its past values, accounting for the correlations at intermediate lags.
* Significance:
* Helps isolate the direct effect of past values on the current value by removing the influence of intermediate values.
* Example: If the bandwidth value 3 hours ago affects the current value, PACF will show this direct correlation, even if the bandwidth value 1 and 2 hours ago also affect the current value.
* Use Case: Useful for identifying the specific lags that have a direct influence on bandwidth, which can help in building more accurate predictive models.

**Summary**

* Mean, Median, Range, and IQR: Provide central tendency and spread of the bandwidth data, indicating typical performance and variability
* Standard Deviation and Coefficient of Variation: Show the variability of the bandwidth data, both in absolute terms and relative to the mean.
* Skewness and Kurtosis: Describe the shape and distribution of the data, indicating asymmetry and the presence of outliers.
* Frequency of Outliers: Highlights unusual bandwidth values that may indicate performance issues.
* ACF and PACF: Analyze the temporal dependencies in the data, identifying patterns, cycles, and direct influences of past values on the current bandwidth.

Summary Statistics Overview

GeorgetownRelay:

Mean (MB/s): 6.97

Standard Deviation (MB/s): 1.81

Median (MB/s): 6.84

IQR (MB/s): 1.92

Skewness: -0.31

Kurtosis: 3.44

Coefficient of Variation: 0.26

Frequency of Outliers: 36

admin (1):

Mean (MB/s): 30.76

Standard Deviation (MB/s): 17.10

Median (MB/s): 36.42

IQR (MB/s): 21.08

Skewness: -1.04

Kurtosis: -0.58

Coefficient of Variation: 0.56

Frequency of Outliers: 0

StuRav:

Mean (MB/s): 29.24

Standard Deviation (MB/s): 3.68

Median (MB/s): 29.78

IQR (MB/s): 5.37

Skewness: -0.14

Kurtosis: -0.69

Coefficient of Variation: 0.13

Frequency of Outliers: 0

Citadel5:

Mean (MB/s): 26.62

Standard Deviation (MB/s): 5.05

Median (MB/s): 27.17

IQR (MB/s): 6.39

Skewness: -0.29

Kurtosis: -0.66

Coefficient of Variation: 0.19

Frequency of Outliers: 0

TORtuga:

Mean (MB/s): 25.95

Standard Deviation (MB/s): 3.31

Median (MB/s): 26.33

IQR (MB/s): 4.97

Skewness: -0.27

Kurtosis: -0.53

Coefficient of Variation: 0.13

Frequency of Outliers: 0

blackfox:

Mean (MB/s): 25.56

Standard Deviation (MB/s): 3.42

Median (MB/s): 25.89

IQR (MB/s): 5.84

Skewness: -0.21

Kurtosis: -0.95

Coefficient of Variation: 0.13

Frequency of Outliers: 0

Unnamed:

Mean (MB/s): 33.86

Standard Deviation (MB/s): 3.43

Median (MB/s): 33.60

IQR (MB/s): 5.13

Skewness: 0.24

Kurtosis: -0.72

Coefficient of Variation: 0.10

Frequency of Outliers: 0

nuker:

Mean (MB/s): 25.67

Standard Deviation (MB/s): 3.33

Median (MB/s): 26.59

IQR (MB/s): 5.55

Skewness: -0.19

Kurtosis: -0.93

Coefficient of Variation: 0.13

Frequency of Outliers: 0

JustExiting:

Mean (MB/s): 8.03

Standard Deviation (MB/s): 2.27

Median (MB/s): 7.80

IQR (MB/s): 2.83

Skewness: -0.18

Kurtosis: 0.55

Coefficient of Variation: 0.28

Frequency of Outliers: 12

**Interpretation and Analysis**

**Standard Deviation:**

Relays like admin (1) and JustExiting exhibit high standard deviations relative to their means, indicating large variability in bandwidth. This suggests these relays experience significant fluctuations, which can impact performance stability.

GeorgetownRelay also shows moderate variability with a noticeable standard deviation.

**Coefficient of Variation (CoV):**

admin (1) has a CoV of 0.56, indicating high variability relative to its mean. This is contrasted with StuRav and Citadel5, which have much lower CoVs (0.13 and 0.19, respectively), indicating more stable performance.

JustExiting has a CoV of 0.28, showing moderate variability.

**Skewness and Kurtosis:**

admin (1) has negative skewness (-1.04) and negative kurtosis (-0.58), suggesting a distribution with a long left tail and fewer extreme high values.

GeorgetownRelay and JustExiting show positive kurtosis, indicating the presence of outliers and a peaked distribution.

Most other relays show negative skewness and kurtosis close to zero, indicating relatively symmetrical and less tailed distributions.

**Frequency of Outliers:**

GeorgetownRelay and JustExiting have a higher frequency of outliers, indicating many extreme values that could impact performance consistency.

Other relays like admin (1), StuRav, and Citadel5 have no outliers in the analyzed period, suggesting more stable performance without extreme fluctuations.

**Conclusion**

Significant Fluctuations: Relays like admin (1) and JustExiting exhibit high variability and sudden fluctuations in bandwidth, as indicated by the high standard deviation and CoV. This supports the argument that timely reaction to performance changes is beneficial.

Moderate Variability: Relays such as GeorgetownRelay and JustExiting show moderate variability with a noticeable number of outliers, indicating periodic performance issues that could benefit from real-time monitoring and adjustment.

Stable Performance: Relays like StuRav, Citadel5, and blackfox demonstrate relatively stable bandwidth performance, with low CoV and no outliers. These relays might not require as frequent adjustments.

Overall, the variability observed in different relays' bandwidths justifies the need for a client-reported method to react to performance changes more timely. This approach can ensure that fluctuations, whether sudden, frequent, or drastic, are promptly addressed to maintain optimal relay performance and network stability.

Make a CDF of the standard dev and CoV?  
Dive into ACF and PACF?

SBWS fluctuations? Can we check for some of these same relays?

Look into connecting with SBWS? How to connect to other relays?

Get most recent consensus file and randomly select 1000’s?

Get lfs?

Migrate Repo’s to GitHub?

Separate big data little data?

Add to repo?

Note how relays are chosen.

Note what consensus file and date.

What defines a fluctuations?

**Key Statistics Calculated**

1. **Mean (MB/s)**:
   * **Definition**: The average bandwidth.
   * **Calculation**: Sum all bandwidth values and divide by the number of values.

Mean=∑Bandwidth ValuesNumber of ValuesMean=Number of Values∑Bandwidth Values​

* + **Significance**: Indicates the typical performance of the relay over the analyzed period.

1. **Standard Deviation (MB/s)**:
   * **Definition**: A measure of the amount of variation or dispersion in the bandwidth values.
   * **Calculation**: Compute the square root of the average of the squared differences from the Mean.

Standard Deviation=∑(Bandwidth Value−Mean)2Number of ValuesStandard Deviation=Number of Values∑(Bandwidth Value−Mean)2​​

* + **Significance**: Shows how much the bandwidth varies from the mean. A higher standard deviation means more variability.

1. **Range (MB/s)**:
   * **Definition**: The difference between the maximum and minimum bandwidth values.
   * **Calculation**: Subtract the minimum bandwidth value from the maximum bandwidth value.

Range=Max Value−Min ValueRange=Max Value−Min Value

* + **Significance**: Provides a sense of the spread of the bandwidth values, indicating the extent of fluctuations.

1. **Coefficient of Variation**:
   * **Definition**: The ratio of the standard deviation to the mean.
   * **Calculation**: Divide the standard deviation by the mean.

Coefficient of Variation=Standard DeviationMeanCoefficient of Variation=MeanStandard Deviation​

* + **Significance**: Indicates the relative variability of the bandwidth. A higher coefficient of variation means more variation relative to the mean. This is a valuable measure because it is relative and helps compare variability across different relays.

1. **Median (MB/s)**:
   * **Definition**: The middle value of the bandwidth data when ordered from lowest to highest.
   * **Calculation**: Sort the bandwidth values and find the middle value.
     + If the number of values is odd, the median is the middle value.
     + If the number of values is even, the median is the average of the two middle values.
   * **Significance**: Provides a measure of central tendency that is less affected by extreme values than the mean.
2. **IQR (Interquartile Range) (MB/s)**:
   * **Definition**: The range between the 25th percentile (Q1) and the 75th percentile (Q3) of the data.
   * **Calculation**: Subtract the 25th percentile value from the 75th percentile value.

IQR=𝑄3−𝑄1IQR=*Q*3−*Q*1

* + **Significance**: Represents the spread of the middle 50% of the data, giving a sense of the typical range of bandwidth values.

1. **Skewness**:
   * **Definition**: A measure of the asymmetry of the data distribution.
   * **Calculation**: Use the formula for skewness, which involves the third central moment.

Skewness=1𝑁∑(Bandwidth Value−Mean)3(1𝑁∑(Bandwidth Value−Mean)2)3/2Skewness=(*N*1​∑(Bandwidth Value−Mean)2)3/2*N*1​∑(Bandwidth Value−Mean)3​

* + **Significance**: Indicates whether the data are skewed to the left (negative skew) or right (positive skew). Symmetrical data has a skewness of zero.

1. **Kurtosis**:
   * **Definition**: A measure of the "tailedness" of the data distribution.
   * **Calculation**: Use the formula for kurtosis, which involves the fourth central moment.

Kurtosis=1𝑁∑(Bandwidth Value−Mean)4(1𝑁∑(Bandwidth Value−Mean)2)2Kurtosis=(*N*1​∑(Bandwidth Value−Mean)2)2*N*1​∑(Bandwidth Value−Mean)4​

* + **Significance**: Indicates the presence of outliers. Higher kurtosis means more outliers.

1. **Frequency of Outliers**:
   * **Definition**: The number of data points that fall outside the typical range (beyond 1.5 times the IQR from Q1 and Q3).
   * **Calculation**:
     + Calculate the IQR.
     + Determine the lower bound: 𝑄1−1.5×𝐼𝑄𝑅*Q*1−1.5×*IQR*
     + Determine the upper bound: 𝑄3+1.5×𝐼𝑄𝑅*Q*3+1.5×*IQR*
     + Count the number of data points below the lower bound or above the upper bound.
   * **Significance**: Highlights unusual bandwidth values, which could indicate performance issues or anomalies.
2. **ACF (Autocorrelation Function)**:
   * **Definition**: Measures the correlation between the current bandwidth value and its past values at different time lags.
   * **Calculation**: Use the formula for autocorrelation, which involves the covariance of the series with a lagged version of itself.

ACF(𝑘)=Cov(Bandwidth(𝑡),Bandwidth(𝑡−𝑘))Var(Bandwidth)ACF(*k*)=Var(Bandwidth)Cov(Bandwidth(*t*),Bandwidth(*t*−*k*))​

* + **Significance**: Helps identify patterns and periodicities in the data.
    - **Example**: If you have bandwidth data recorded every hour, ACF will show how similar the bandwidth is to its values from 1 hour ago, 2 hours ago, etc.
    - **Use Case**: Useful for detecting regular fluctuations or cycles in bandwidth, which could indicate recurring performance issues.

1. **PACF (Partial Autocorrelation Function)**:
   * **Definition**: Measures the correlation between the current bandwidth value and its past values, accounting for the correlations at intermediate lags.
   * **Calculation**: Use the formula for partial autocorrelation, which adjusts the autocorrelation to remove the influence of intermediate values.

PACF(𝑘)=Autocorrelation(𝑘)−Linear combination of autocorrelations at lags 1 to 𝑘−1PACF(*k*)=Autocorrelation(*k*)−Linear combination of autocorrelations at lags 1 to *k*−1

* + **Significance**:
    - **Example**: If the bandwidth value 3 hours ago affects the current value, PACF will show this direct correlation, even if the bandwidth value 1 and 2 hours ago also affect the current value.
    - **Use Case**: Useful for identifying the specific lags that have a direct influence on bandwidth, which can help in building more accurate predictive models.

**Understanding Lags**

1. **Lag 1**:
   * **Definition**: The value of the data series from one time step before the current time step.
   * **Example**: If today is 𝑡*t*, the value at lag 1 is the value from 𝑡−1*t*−1.
2. **Lag 2**:
   * **Definition**: The value of the data series from two time steps before the current time step.
   * **Example**: If today is 𝑡*t*, the value at lag 2 is the value from 𝑡−2*t*−2.
3. **Lag k**:
   * **Definition**: The value of the data series from 𝑘*k* time steps before the current time step.
   * **Example**: If today is 𝑡*t*, the value at lag 𝑘*k* is the value from 𝑡−𝑘*t*−*k*.

**Importance of Lags in ACF and PACF**

1. **Autocorrelation Function (ACF)**:
   * **Purpose**: Measures how the current value of the series is related to its previous values (lags).
   * **Calculation**: Correlates the data series with lagged versions of itself.
   * **Use**: Helps identify patterns such as seasonality or trends by showing how past values influence current values at different lags.
2. **Partial Autocorrelation Function (PACF)**:
   * **Purpose**: Measures the correlation between the current value and lagged values, controlling for the influence of intermediate lags.
   * **Calculation**: Adjusts the correlation to isolate the direct effect of a specific lag.
   * **Use**: Identifies the direct relationship between the current value and past values, useful for understanding the intrinsic properties of the data series and building predictive models.

**Examples**

* **ACF Example**:
  + If you have bandwidth data recorded every hour, ACF at lag 1 shows the correlation between the bandwidth at hour 𝑡*t* and hour 𝑡−1*t*−1.
  + ACF at lag 2 shows the correlation between the bandwidth at hour 𝑡*t* and hour 𝑡−2*t*−2, and so on.
* **PACF Example**:
  + PACF at lag 1 shows the direct correlation between the bandwidth at hour 𝑡*t* and hour 𝑡−1*t*−1, ignoring intermediate values.
  + PACF at lag 2 shows the direct correlation between the bandwidth at hour 𝑡*t* and hour 𝑡−2*t*−2, after removing the effect of the bandwidth at hour 𝑡−1*t*−1.

**Use Case in Bandwidth Analysis**

* **Detecting Patterns**:
  + **ACF** can reveal repeating patterns, such as daily or weekly cycles in bandwidth usage, by showing how strongly the bandwidth at one time point is related to bandwidth at previous time points.
  + **PACF** helps in identifying specific past bandwidth values that have a direct effect on the current bandwidth, crucial for understanding causal relationships and building time series models.
* **Model Building**:
  + Knowing the lags that significantly influence current bandwidth can help in constructing autoregressive models (AR models), where future values are predicted based on past values.

**Practical Steps**

1. **Calculate ACF and PACF**:
   * Use functions from statistical libraries (e.g., **acf** and **pacf** from **statsmodels** in Python) to compute these values for your data.
2. **Interpret the Plots**:
   * **ACF Plot**: Look for lags where the correlation is significantly high. Peaks at regular intervals might indicate a seasonal effect.
   * **PACF Plot**: Identify the significant lags, which help in selecting the order of autoregressive models.

Understanding lags and their role in ACF and PACF is critical for analyzing time series data, identifying patterns, and building predictive models that account for the temporal dependencies in the data.