README

Project Overview

This project aims to analyze the bandwidth fluctuations of Tor network relays by collecting and processing bandwidth data. The goal is to understand how relay bandwidth varies over time and to interpret these variations to improve relay performance monitoring. This analysis helps in identifying the stability and reliability of the relays and can support the new approach where clients report throughput they observe, providing timely performance feedback.

Features

- **Fetch Bandwidth Data**: Collects bandwidth data for specified relays from the Tor network using the Onionoo API.
- **Data Processing**: Processes the collected data to extract recent bandwidth usage.
- **Statistical Analysis**: Calculates various statistics to interpret bandwidth fluctuations.
- **Data Visualization**: Generates plots to visualize bandwidth over time, histogram of bandwidth values, scatter plots, and autocorrelation and partial autocorrelation plots.
- **Excel Report**: Saves the statistical analysis and visualizations to an Excel file.

Statistics Calculated

- **Mean (MB/s)**: Average bandwidth usage over the specified period.
- **Standard Deviation (MB/s)**: Measure of bandwidth variation or dispersion from the mean.
- **Range (MB/s)**: Difference between maximum and minimum bandwidth values.
- **Coefficient of Variation**: Ratio of the standard deviation to the mean, indicating relative variability.
- **Median (MB/s)**: Middle value of the bandwidth data, providing a robust measure of central tendency.
- **IQR (MB/s)**: Interquartile range, indicating the range within which the middle 50% of the data lies.
- **Skewness**: Measures the asymmetry of the bandwidth distribution.
- **Kurtosis**: Measures the "tailedness" of the bandwidth distribution.
- **ACF**: Autocorrelation function values, showing the correlation of bandwidth values with lagged values.
- **PACF**: Partial autocorrelation function values, showing the correlation of bandwidth values with lagged values, controlling for intermediate lags.

How to Run

1. **Install Required Libraries**:

```
pip install pandas numpy matplotlib openpyxl requests statsmodels scipy
```

2. **Run the Script**:

```
python analyze_bandwidth.py <input_excel_filename>
```

- `<input_excel_filename>`: Path to the input Excel file containing relay fingerprints and names.

Input Excel File Format

The input Excel file should have the following columns:

- **Relay Name**: Name of the relay.
- **Fingerprint**: Fingerprint of the relay.

Output

- **Excel Report**: The script generates an Excel file `Relays_Analysis.xlsx` containing the following:
  - **Relay Sheets**: Each relay has its own sheet with statistical analysis and visualizations.
  - **Summary Sheet**: Includes aggregated statistics and bar plots for mean, standard deviation, median, IQR, skewness, kurtosis, and coefficient of variation across all relays.

Detailed Explanation of Statistics

Mean (MB/s)

- **Description**: The average bandwidth usage over the specified period.
- **Purpose**: Indicates consistent handling of data by the relay. High mean signifies high data handling capacity.

Standard Deviation (MB/s)

- **Description**: Measures the amount of variation or dispersion of bandwidth values from the mean.
- **Purpose**: Low standard deviation indicates stable bandwidth usage, while high standard deviation suggests significant fluctuations.

Range (MB/s)

- **Description**: Difference between the maximum and minimum bandwidth values recorded.
- **Purpose**: Indicates the extremes of bandwidth usage, highlighting periods of heavy usage or inactivity.

Coefficient of Variation

- **Description**: Ratio of the standard deviation to the mean.
- **Purpose**: Provides a normalized measure of bandwidth variability. High coefficient indicates inconsistent usage.

Median (MB/s)

- **Description**: The middle value of the bandwidth data.
- **Purpose**: Offers a robust measure of central tendency, especially useful in the presence of outliers.

Interquartile Range (IQR)

- **Description**: The range within which the middle 50% of the data lies.
- **Purpose**: Helps understand the spread of the central portion of the data, providing insight into typical variability.

Skewness

- **Description**: Measures the asymmetry of the distribution of bandwidth values.
- **Purpose**: Indicates whether the bandwidth data is skewed towards higher or lower values.

Kurtosis

- **Description**: Measures the "tailedness" of the bandwidth distribution.
- **Purpose**: Helps identify the presence of outliers and the propensity for extreme values.

Autocorrelation (ACF)

- **Description**: Measures how current bandwidth values are related to past values.
- **Purpose**: Detects periodic patterns or trends in bandwidth usage.

Partial Autocorrelation (PACF)

- **Description**: Measures the correlation between bandwidth values and their lagged values, controlling for intermediate lags.
- **Purpose**: Helps identify direct relationships in the time series data.

Aggregating Data Among Relays

Aggregating the data involves summarizing the statistics for all relays to understand overall trends and variability. This helps in identifying relays that significantly deviate from the norm and in understanding the general behavior of the network.

Aggregated Statistics

- **Mean of Means**: The average of the mean bandwidths of all relays. It provides an overall average bandwidth usage.
- **Mean of Standard Deviations**: The average of the standard deviations of all relays. It provides an overall measure of bandwidth variability.
- **Overall Range**: The difference between the maximum and minimum bandwidth values across all relays.

Interpretation:

- By comparing individual relay statistics to the aggregated statistics, we can identify relays that are outliers or that have unusual behavior.
- Aggregated statistics help in understanding the general performance and reliability of the network.

Visualization for Better Interpretation

Time Series Plot

A time series plot shows the bandwidth usage over time, helping to identify trends, seasonal patterns, and anomalies.

Histogram

A histogram shows the distribution of bandwidth values, helping to understand the frequency of different bandwidth levels.

Scatter Plot

A scatter plot of bandwidth values against time helps in visualizing the spread and clustering of data points, indicating periods of high or low activity.

Statistics Explanation

Mean

Mean (MB/s): The mean bandwidth represents the average bandwidth usage over the specified period. It gives a general idea of the typical bandwidth the relay is handling.

Significance: High mean bandwidth indicates the relay is consistently handling large amounts of data, while a low mean indicates lower usage.

Standard Deviation

Standard Deviation (MB/s): The standard deviation measures the amount of variation or dispersion from the mean. A high standard deviation indicates that the bandwidth usage fluctuates greatly, while a low standard deviation indicates more stable bandwidth usage.

Significance: Understanding the standard deviation helps in assessing the reliability of the relay's bandwidth. Large fluctuations might imply periods of congestion or underutilization.

Range

Range (MB/s): The range is the difference between the maximum and minimum bandwidth values recorded. It provides an insight into the extremes of the relay's bandwidth usage.

Significance: A large range indicates that the relay experiences significant peaks and troughs in bandwidth usage. This can help identify if there are times of high demand or underuse.

Coefficient of Variation

Coefficient of Variation: This is the ratio of the standard deviation to the mean. It provides a normalized measure of dispersion of the bandwidth data.

Significance: The coefficient of variation helps compare the degree of variation from one relay to another, irrespective of their means. A high coefficient indicates a high level of fluctuation relative to the mean.

Understanding the Metrics

Interpreting Computed Values

Mean (MB/s)

Mean (MB/s): The mean bandwidth represents the average bandwidth usage over the specified period. It is calculated by summing all the bandwidth measurements and dividing by the number of measurements.

Interpretation:

- **High Mean**: Indicates that the relay is consistently handling a high amount of data. This could be desirable for high-traffic relays.
- **Low Mean**: Indicates lower usage, which could be due to fewer users or less activity on the network.

Standard Deviation (MB/s)

Standard Deviation (MB/s): The standard deviation measures the amount of variation or dispersion of bandwidth values from the mean. It is calculated as the square root of the variance, where variance is the average of the squared differences from the mean.

Interpretation:

- **Low Standard Deviation**: Indicates that the bandwidth usage is stable and does not fluctuate much from the mean. This is desirable for consistency and predictability.
- **High Standard Deviation**: Indicates significant fluctuations in bandwidth usage, suggesting periods of congestion or underutilization.

Example:

- A standard deviation of 1 MB/s with a mean of 5 MB/s means that most of the bandwidth values fall within 4 to 6 MB/s.
- A standard deviation of 3 MB/s with the same mean indicates a wider spread, with bandwidth values ranging from 2 to 8 MB/s.

Range (MB/s)

Range (MB/s): The range is the difference between the maximum and minimum bandwidth values recorded. It provides insight into the extremes of the relay's bandwidth usage.

Interpretation:

- **Large Range**: Indicates that the relay experiences significant peaks and troughs in bandwidth usage. This could suggest occasional heavy usage or very low activity periods.
- **Small Range**: Indicates that the bandwidth usage is more consistent without extreme values.

Coefficient of Variation

Coefficient of Variation: This is the ratio of the standard deviation to the mean. It provides a normalized measure of dispersion of the bandwidth data.

Interpretation:

- **High Coefficient of Variation**: Indicates a high level of fluctuation relative to the mean. This could mean that the relay's bandwidth usage is very inconsistent.
- **Low Coefficient of Variation**: Indicates low variability relative to the mean, suggesting more consistent bandwidth usage.

Additional Statistics for Interpretation

Skewness

Skewness measures the asymmetry of the distribution of bandwidth values. It helps understand whether the data are skewed to the left (negative skew) or to the right (positive skew).

Interpretation:

- **Negative Skewness**: Indicates that the left tail is longer or fatter than the right. Most of the values are concentrated on the higher end.
- **Positive Skewness**: Indicates that the right tail is longer or fatter than the left. Most of the values are concentrated on the lower end.
- **Zero Skewness**: Indicates a symmetric distribution.

Kurtosis

Kurtosis measures the "tailedness" of the bandwidth distribution. It helps identify the presence of outliers.

Interpretation:

- **High Kurtosis**: Indicates a distribution with heavy tails or outliers.
- **Low Kurtosis**: Indicates a distribution with light tails, suggesting fewer outliers.

Understanding the Values

Standard Deviation Interpretation

- **Low Standard Deviation**: Indicates that the relay's bandwidth is stable and does not fluctuate much from the mean. This is desirable for consistency.
- **High Standard Deviation**: Suggests significant fluctuations in bandwidth, which might indicate periods of congestion or underutilization.

Aggregating Data Among Relays

Aggregating the data involves summarizing the statistics for all relays to understand overall trends and variability. This helps in identifying relays that significantly deviate from the norm and in understanding the general behavior of the network.

Aggregated Statistics

- **Mean of Means**: The average of the mean bandwidths of all relays. It provides an overall average bandwidth usage.
- **Mean of Standard Deviations**: The average of the standard deviations of all relays. It provides an overall measure of bandwidth variability.
- **Overall Range**: The difference between the maximum and minimum bandwidth values across all relays.

Interpretation:

- By comparing individual relay statistics to the aggregated statistics, we can identify relays that are outliers or that have unusual behavior.

- Aggregated statistics help in understanding the general performance and reliability of the network.

Conclusion

By analyzing these statistics and visualizations, we can gain a comprehensive understanding of how much and in what way the relays' bandwidth fluctuates. This helps in assessing the reliability and performance of the relays and in identifying potential issues or areas for improvement.