Relay Bandwidth Analysis

This project is designed to analyze the bandwidth history of Tor relays. The script fetches bandwidth data, computes various statistics, generates plots, and outputs the results into an Excel file with each relay's data in its own sheet and a summary sheet with aggregated statistics.

Table of Contents

Installation

1. Clone this repository or download the script `analyze_bandwidth.py`.

2. Ensure you have Python 3.6 or later installed.

3. Install the required packages by running:

   ```
   pip install -r requirements.txt
   ```

   The `requirements.txt` should contain:

   ```
   requests
   pandas
   matplotlib
   openpyxl
   ```

   ## Usage

4. Prepare an Excel file (e.g., `RelayList.xlsx`) with two columns: "Relay Name" and "Fingerprint".

5. Run the script:

   ```
   python analyze_bandwidth.py RelayList.xlsx
   ```

6. The script will generate an output file named `Relays_Analysis.xlsx`.

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 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.

Other Statistics

- **Mean**: Helps in understanding the average bandwidth usage, indicating the typical load the relay handles.

- **Range**: Useful for identifying extreme values in bandwidth usage, which might indicate occasional heavy usage or very low activity periods.
- **Coefficient of Variation**: Provides insight into the relative variability of the bandwidth, making it easier to compare different relays.

Aggregating Data Among Relays

Aggregating the data involves looking at the overall mean, standard deviation, range, and coefficient of variation across multiple relays. This helps in understanding general trends and identifying relays that significantly deviate from the norm. The summary sheet in the output Excel file provides plots for the total mean and standard deviation of each relay, making it easier to visualize and interpret the data.

File Descriptions

- **Relay.py**: The main script to analyze relay bandwidth data.
- **RelayList.xlsx**: Example input file with relay names and fingerprints.
- **Relays_Analysis.xlsx**: The output file containing detailed analysis and plots for each relay.

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.

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.

Conclusion

By interpreting 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.