Multi-Level Anomaly Detection for Correlated Network Traffic Patterns

A Comprehensive Approach to Network Traffic Analysis

A Technical Report By Nebojsa Djosic, Evgenii Ostanin, Salah Sharieh
Toronto, ON, May 2025

Abstract

This paper addresses the complex challenge of anomaly detection in network traffic data for remote (egress) destinations, where traditional single-variable approaches fail to capture the intricate relationships and causation patterns between different services. We propose a multi-level detection framework that combines individual remote (egress) destination modeling with correlation-aware multivariate analysis to provide comprehensive anomaly detection across daily, weekly, and monthly timeframes.

1. Problem Definition

1.1 Data Characteristics

The dataset under analysis consists of network traffic byte counts collected hourly over a multi-month historical period, with ongoing daily updates. The data structure includes:

- **Destination Name (dest)**: Over 350 unique egress remote destinations, each representing a set of remote (target) IPs associated with an identifiable, recognized service, such as Box, Slack, WebEx, etc.
- **Temporal resolution**: Measurements are collected in daily files from midnight to midnight, and aggregated hourly in the format `yyyymmddhh`
- **Traffic metrics**: `bytes_in` and `bytes_out` counts per destination per hour for 24 hours for each calendar day

- **Feature engineering requirements**: Total bytes, byte ratios, percentage distributions, z-scores, and other statistics are added during pre-processing to create datasets in CSV file format.

1.2 Traffic Pattern Complexity

We identified many complex patterns in the data that need special attention and handling to be addressed. In this analysis, we are focusing on two distinct behavioural patterns among destinations:

Pattern A - Spikes with Stable Ratios: Destination data shows natural, random traffic spikes in both inbound and outbound traffic. For instance, a large file is uploaded to S3, or a data-hungry application runs for a period of time following an event. The traffic observed would resemble a heart rate with or without regularity in both frequency and amplitude. However, this pattern would maintain relatively stable byte ratios (in/out), percentages, z-scores, and other statistical data, with some minor fluctuation.

Pattern B - Time-of-day-dependent: Data is displaying predictable temporal patterns with steady increases from 06:00-10:00 AM, relatively minor peak fluctuations from 10:00 AM-6:00 PM, followed by a gradual decline until 10:00 PM, and low-volume periods typically from 10:00 PM-06:00 AM.

1.3 Correlation and Causation Challenges

While we identified the need to have distinct anomaly detection per destination, we also noticed complex correlation and causation that have to be taken into account and addressed. The primary complexity arises from inter-destination relationships where some destinations exhibit strong correlations, and causation effects manifest as compensatory behaviours—when one service experiences reduced traffic, correlated services may increase, or groups of services may exhibit synchronized patterns.

2. Literature Review and Current State

Recent research in network traffic anomaly detection has increasingly focused on multivariate approaches. Machine learning offers powerful techniques to accurately detect anomalies by analyzing patterns in network data, with anomaly detection helping prevent cyber threats and ensure network reliability by identifying unusual traffic patterns that may signal security breaches or unauthorized access.

Due to the complexity of inconspicuous anomalies, high dynamicity, and the lack of anomaly labels in the cloud environment, multivariate time series anomaly detection becomes more difficult, with existing approaches rarely effective in meeting these challenges. This complexity is particularly evident in cloud environments, where few studies have approached the intrusion detection problem as a time series issue, requiring time series modelling.

The challenge of multivariate time series anomaly detection is well-documented, as MTSAD is challenging due to the need for simultaneous analysis of temporal and spatial dependencies. Research shows that performing anomaly detection on multivariate time series data can timely find faults, prevent malicious attacks, and ensure safe and reliable operation, however, the rarity of abnormal instances leads to a lack of labeled data.

3. Proposed Multi-Level Detection Framework

3.1 Architectural Overview

Our approach implements a hierarchical three-level detection system that addresses both individual destination anomalies and inter-destination relationship disruptions:

Level 1: Individual destination anomaly detection (univariate analysis)

Level 2: Correlation group anomaly detection (multivariate within groups)

Level 3: System-wide anomaly detection (global multivariate analysis)

3.2 Level 1: Individual Destination Modeling

3.2.1 Pattern-Specific Approaches

For **Stable Ratio Destinations**:

- Implement ratio deviation monitoring with adaptive thresholds (to account for “natural” spikes)
- Apply volume-based anomaly detection using z-score (and other statistical) analysis
- Monitor for sudden ratio instability, indicating potential data exfiltration or service disruption

For **Time-Dependent Destinations**:

- Utilize time-series decomposition (trend, seasonal, residual components)
- Implement seasonal AutoRegressive Integrated Moving Average (ARIMA)***** modelling for expected pattern prediction. ARIMA only indirectly detects anomalies and it’s univariate only, but it is the best for *predicting* daily bytes transferred to a remote destination, allowing also for proactive monitoring. Statistical autoregressive (AR) models such as ARIMA, Vector Autoregression (VAR), VARIMA have also been extensively studied in the past for detecting outliers in both univariate and multivariate time series, but they work through prediction accuracy rather than direct pattern recognition. In our case, given the requirement for proactive monitoring, this approach provides for both anomaly detection and prediction. Given this complexity, we suggest a hybrid approach:
    - ARIMA for individual stable-ratio destinations - captures their specific patterns
    - VAR for correlated groups - handles the causation relationships you mentioned

    - Ensemble residual analysis - combines both approaches' prediction errors
- Apply adaptive baseline learning to accommodate gradual pattern evolution

3.2.2 Feature Engineering Enhancement

Beyond basic metrics, implement advanced feature extraction:

- Rolling window statistics (mean, variance, skewness, kurtosis)
- Time-based features (hour-of-day, day-of-week, month effects)
- Rate-of-change indicators for trend analysis
- Frequency domain features using FFT for periodic pattern detection

3.3 Level 2: Correlation Group Analysis

3.3.1 Correlation Discovery

Recent advances in multivariate analysis show that topology-aware multivariate time series anomaly detectors use GNN, LSTM, and VAE for spatiotemporal learning, leveraging topological information to identify anomalies using graph-based approaches.

Implement dynamic correlation clustering:

- Sliding window Pearson correlation analysis, used to measure the strength and direction of the linear relationship between two continuous variables, and is good for discovering unexpected dependencies, but it's linear only, sensitive to outliers, and assumes continuous data
- Hierarchical clustering (dendrogram) based on (pairwise) correlation coefficients is a technique to group observations (rows, hourly metrics) by their similarity, specifically by how correlated they are with one another.
- Dynamic (daily, weekly, monthly, annual) reclustering to adapt to changing relationship patterns

3.3.2 Causation Analysis

Granger causality testing is a statistical method used to determine whether one *time series* can help predict another. Despite the name, it does not prove true causality — it checks if the past values of one variable provide statistically significant information about the future values of another. "Variable X Granger-causes Y if past values of X help predict Y beyond what Y's own past can do." Granger causality is not true causality — it shows predictive precedence. Also, outliers and autocorrelation can distort results. We assume that our data is a *stationary time series*, one whose statistical properties do not change over time. Clearly, we must ensure that the following is true over the time periods we use, which means that the data must be carefully preprocessed, analyzed and selected before we apply the analysis:

1. *Constant mean*: the average values don't drift (increase or decrease) over the time periods we use.

2. *Constant variance*: the spread (or volatility) remains consistent over the time periods we use.
3. *Constant autocorrelation*: the relationship between past and present values is stable over the time periods we use.

We are going to apply Granger causality testing to distinguish true causal relationships from spurious correlations:

- Sliding window Granger causality tests with varying lag parameters
- Transfer entropy analysis for non-linear causal relationships
- Anomaly detection when established causal relationships weaken or strengthen unexpectedly

3.3.3 Group-Level Anomaly Detection

For identified correlation groups:

- Principal Component Analysis (PCA) for dimensional reduction
- Multivariate control charts like Hotelling's T^2 statistic, a measure of how far a multivariate row is from the center (mean vector) of a reference distribution, accounting for the correlation between variables, or a multivariate generalization of the z-score. The advantage is that it detects joint anomalies, not just marginal ones
- Isolation Forest applied to group-level feature spaces
- Long Short-Term Memory (LSTM) Autoencoder networks for complex temporal pattern learning in sequential data will be trained on normal behaviour only to accurately reconstruct normal patterns. During real-time testing, when given anomalous data, the reconstruction error becomes high because the model hasn't seen it before.

3.4 Level 3: System-Wide Analysis

3.4.1 Global Anomaly Detection

Implement system-wide monitoring using:

- Global traffic volume anomalies across all destinations
- Network-wide ratio distributions and their deviations
- Cross-correlation matrix anomalies indicating relationship disruptions
- Entropy-based measures for traffic distribution normality

3.4.2 Ensemble Approach

Combine multiple detection algorithms:

- Statistical methods like z-score and interquartile range (IQR) based detection. IQR is not affected by extreme values like mean/std, works with non-Gaussian or skewed data, is easy to explain and implement, and is ideal for univariate anomaly detection.

- Machine learning approaches like Isolation Forest and/or One-Class SVM.
- Deep learning methods like Variational Autoencoders and/or LSTM-Autoencoders
- Time-series specific methods like ARIMA residuals

4. Implementation Framework

4.1 Multi-Temporal Analysis

The framework supports anomaly detection across multiple time horizons:

Daily Analysis:

- Real-time processing of the previous day's 24-hour data
- Immediate alerting for critical anomalies
- Adaptive threshold updates based on recent patterns

Weekly Analysis:

- Pattern consistency evaluation across weekly cycles
- Detection of weekly trend anomalies
- Seasonal adjustment for business cycle effects

Monthly and Yearly Analysis:

- Long-term trend anomaly identification
- Baseline model retraining and validation
- Strategic pattern evolution analysis

4.2 Adaptive Learning System

The system incorporates continuous learning mechanisms:

- Online learning for threshold adaptation
- Incremental model updates to reduce computational overhead
- Feedback incorporation for false positive reduction *requiring human-in-the-middle*
- Drift detection for model relevance maintenance

5. Expected Outcomes and Applications

This multi-level approach addresses key limitations in current anomaly detection systems typically in use by:

1. **Reducing False Positives**: By considering inter-destination relationships, the system can distinguish between coordinated normal behaviour and true anomalies

2. **Improving Detection Sensitivity**: Multi-level analysis captures subtle anomalies that might be missed by single-variable approaches
3. **Providing Contextual Understanding**: Correlation and causation analysis offer insights into why anomalies occur
4. **Supporting Multiple Timeframes**: Daily, weekly, and monthly analysis provides comprehensive temporal coverage

6. Conclusion

The proposed multi-level anomaly detection framework addresses the complex challenge of identifying anomalies in correlated cloud service traffic patterns. By combining individual destination modelling with sophisticated correlation and causation analysis, this approach provides a comprehensive solution for network traffic anomaly detection that considers both temporal patterns and inter-service relationships.

The framework's hierarchical structure enables scalable implementation while maintaining detection accuracy, making it suitable for enterprise-scale network monitoring applications where understanding traffic pattern relationships is crucial for security and operational efficiency.

---

* Moayedi, H. & Masnadi-Shirazi, M.A.. (2008). ARIMA model for network traffic prediction and anomaly detection. Proceedings of the International Symposium on Information Technology. 4. 1 - 6.

> 10.1109/ITSIM.2008.4631947. *This paper presents the use of a basic ARIMA model for network traffic prediction and anomaly detection. Accurate network traffic modelling and prediction are important for network provisioning and problem diagnosis, but network traffic is highly dynamic. To achieve better modelling and prediction, it is needed to isolate anomalies from normal traffic variation. Thus, we decompose traffic signals into two parts: normal variations, that follow certain law and are predictable and, anomalies that consist of sudden changes and are not predictable. ARIMA analysis and modelling for network traffic prediction is able to detect and identify volume anomaly or outliers.*

References

1. Zhang, Y., et al. (2023). "Anomaly detection using spatial and temporal information in multivariate time series." *Scientific Reports*, Nature Publishing Group. https://www.nature.com/articles/s41598-023-31193-8

2. Eyer AI. (2024). "Network Traffic Anomaly Detection with Machine Learning." *AI Technology Blog*. https://www.eyer.ai/blog/network-traffic-anomaly-detection-with-machine-learning/

3. Chen, L., et al. (2023). "A Novel Convolutional Adversarial Framework for Multivariate Time Series Anomaly Detection and Explanation in Cloud Environment." *Applied Sciences*, MDPI. https://www.mdpi.com/2076-3417/12/20/10390

4. Liu, H., et al. (2023). "Intrusion detection in cloud computing based on time series anomalies utilizing machine learning." *Journal of Cloud Computing*, Springer. https://journalofcloudcomputing.springeropen.com/articles/10.1186/s13677-023-00491-x

5. Kumar, S., et al. (2024). "Machine Learning and Deep Learning Techniques for Internet of Things Network Anomaly Detection—Current Research Trends." *Sensors*, MDPI. https://www.mdpi.com/1424-8220/24/6/1968

6. Wang, X., et al. (2024). "Machine Learning-Based Network Anomaly Detection: Design, Implementation, and Evaluation." *Journal of Network and Computer Applications*. https://www.mdpi.com/2673-2688/5/4/143

7. He, K., et al. (2020). "Deep Learning for Time Series Anomaly Detection: A Survey." *arXiv preprint*. https://arxiv.org/abs/2211.05244

Other sources:

8. Anomaly Detection in Time Series
9. Univariate Time Series Anomaly Detection Using ARIMA Model
10. ARIMA model for network traffic prediction and anomaly detection
11. Online Forecasting and Anomaly Detection Based on the ARIMA Model
12. ARIMA based algorithms vs neural networks in anomaly detection
13. Multivariate Time Series Anomaly Detection using VAR model
14. Anomaly Detection in Multivariate Time Series with VAR | Towards Data Science
15. Anomaly detection using vector autoregression - Cross Validated
16. Unsupervised Anomaly Detection for IoT-Based Multivariate Time Series: Existing Solutions, Performance Analysis and Future Directions