جامعة الأخوين

AL AKHAWAYN
UNIVERSITY

School of Science and Engineering

Engineering Stochastic Processes (EGR4394)

Project Report

Handling Uncertainty and Risks in Network Security

Spring 2024

El Mehdi Ziate

CONTENT

LIST OF FIGURES

ABSTRACT (ENGLISH)

With the increasing sophistication and the dynamic increase of vulnerabilities, cyber treats become a serious talk nowadays, and the nature of dynamic nature of network traffic, system compromises often occur unpredictably which complicates detection efforts. This project leverages advanced stochastic modeling techniques to analyze network traffic data, aiming to detect system compromises through the identification of anomalies. The data, sourced from Kaggle and originally collected by Stanford University, spans a three-month period and records daily network activity for 10 local workstation IPs, half of which were compromised. By employing real-time threat intelligence and dynamic risk assessment strategies, the project seeks to enhance cybersecurity measures through data-driven insights.

Keywords: (Network, Stochastics, Models, Cybersecurity)

RESUME (FRENCH)

Avec l'augmentation de la sophistication et la multiplication dynamique des vulnérabilités, les menaces cybernétiques sont devenues un sujet sérieux de nos jours. De plus, en raison de la nature dynamique du trafic réseau, les compromissions de systèmes surviennent souvent de manière imprévisible, ce qui complique les efforts de détection. Ce projet utilise des techniques de modélisation stochastique avancées pour analyser les données de trafic réseau, dans le but de détecter les compromissions de systèmes à travers l'identification d'anomalies. Les données, provenant de Kaggle et initialement collectées par l'Université de Stanford, couvrent une période de trois mois et enregistrent l'activité réseau quotidienne de 10 adresses IP de postes de travail locaux, dont la moitié a été compromise. En employant des stratégies de renseignement sur les menaces en temps réel et d'évaluation dynamique des risques, le projet vise à renforcer les mesures de cybersécurité grâce à des analyses basées sur les données.

Mots clés: Réseau, Stochastique, Modèles, Cybersécurité

1. INTRODUCTION: PROJECT IDENTIFICATION

1.1. INTRODUCTION

In the context of cybersecurity and complex system of network security, system compromises emerge as a challenge fraught with uncertainty and randomness. The dynamics of network traffic, characterized by a continuous stream of data flowing between interconnected devices and systems, is a fertile sole for spreading malicious activities.

System compromises frequently follow erratic patterns due to the variety of vulnerabilities present in network infrastructures and the constantly changing strategies used by cybercriminals. The difficulty is compounded by the vast volume and diversity of network traffic, which masks the obvious indicators of compromise from view behind the cacophony of official communications. Because of this, identifying and addressing system breaches requires a deep comprehension of the random character of cyberattacks in addition to advanced analytical methods that can identify minute irregularities in the midst of a flurry of network activity.

In this milestone report, we outline a project aimed at detecting system compromises in a network environment by analyzing network traffic data. We focus on handling uncertainty, risks, and dynamic decision problems inherent in network security. The project utilizes a dataset containing daily records of network traffic for different local workstation IPs over a three-month period. The dataset includes indications of compromised systems, enabling us to explore anomalies in network traffic that can indicate a compromise.

1.2. HIGHLIGHT OF UNCERTAINTY, RISKS AND DYNAMIC DECISION

1.2.1. PROJECT SELECTION AND CONTEXT

In selecting a project for handling uncertainty and risks in network security, we recognize the multifaceted challenges inherent in protecting digital assets from cyber threats. Network security operates within an environment characterized by continuous changes and uncertainties. The project aims to address these complexities by leveraging data-driven approaches to detect and mitigate potential security breaches in a dynamic network landscape. By focusing on anomaly detection through network traffic analysis, the project confronts the dynamic nature of cyber threats and the uncertainty in distinguishing between benign and malicious activities.

1.2.2. UNCERTAINTY

The uncertainty in network security arises from the variable patterns of network traffic and the sophisticated methods employed by cyber adversaries. The dynamic nature of network traffic, coupled with the evolving tactics of cyber adversaries, introduces inherent unpredictability. For instance, anomalies in network traffic may signify benign activity or indicate a sophisticated cyber-attack, making it challenging to differentiate between normal and malicious behavior. These challenges are met with a stochastic modeling approach, enabling us to deal with the inherent randomness and to estimate the probability of system compromises, rather than making definitive predictions.

1.2.3. RISK

The risks associated with network security breaches are manifold and parallel to the risks observed in currency exchange fluctuations. These risks manifest at different levels, impacting organizations, economies, and individuals:

- **Operational Risks**: We consider the potential disruptions and damage caused by security breaches, including data theft and service downtime.

- **Economic Risks**: The broader economic implications of network security breaches are examined, noting their potential to impact productivity and trust in digital infrastructure.

- **Regulatory Risks**: We consider the compliance and legal ramifications of network security, including fines and mandatory disclosures.

- **Investment Risks**: The necessity of investments in cybersecurity measures is discussed, along with the risks associated with inadequate security protocols.

1.2.4. DYNAMIC DECISION CONTEXT

The dynamic decision-making in network security is characterized by constant vigilance and adaptability to counteract evolving cyber threats. Key elements include:

- **Real-time Threat Intelligence**: Security teams use up-to-the-minute information to promptly identify and respond to new cyber threats, ensuring rapid incident detection and response through continuous network surveillance and system monitoring.

- **Dynamic Risk Assessment**: Ongoing evaluation of system vulnerabilities and potential threats allows decision-makers to modify risk mitigation strategies in response to resources.

- **Incident Response and Recovery**: Organizations implement responsive and flexible incident response plans to quickly address security breaches, aiming to reduce the impact, recover operations swiftly, and prevent significant downtime or data loss.

- **Adaptive Security Measures**: Security measures are constantly refined in light of new threats and user behaviors, with updates to security settings, access controls, and detection algorithms to maintain robust defenses against cyber risks.

2. PROJECT DESCRIPTION

2.1. GOALS & OBJECTIVES OF THE PROJECT

The primary goal of this project is to develop and implement a robust stochastic model that can effectively analyze network traffic data to predict daily connection flows. This predictive model will classify the traffic flow as either 'low' or 'high,' which will serve as indicators of potential system compromises. By accomplishing this, the project seeks to enhance the capabilities of cybersecurity systems in detecting and responding to anomalies that may signify security breaches.

2.2. BENCHMARKING VIS-À-VIS CURRENT STATE OF THE ART

Comparative to the SISTER model discussed in Silva et al.'s paper (Silva et al., 2021), which utilizes Stochastic Activity Networks to handle uncertainties in tramway systems, our proposed use of a Discrete-Time Markov Chain offers a targeted approach to handle network traffic dynamics for cybersecurity. While both models apply stochastic principles, the DTMC is specifically aligned with the categorical and sequential nature of network data, focusing on state transitions within network traffic to detect anomalies effectively. This makes our approach particularly suited to analyzing and predicting system compromises in dynamic network environments, thereby enhancing predictive accuracy and system resilience against cyber threats.

3. DATA COLLECTION AND EXPLORATION

3.1. METHODS OF DATA COLLECTION

For the project on "Handling Uncertainty and Risks in Network Security," we employed an open-source dataset from Kaggle to analyze network traffic. This dataset, named "Computer Network Traffic," originates from Stanford University's StatWeb and is publicly accessible on the

Kaggle platform. Utilizing this dataset enables a thorough examination of network anomalies and potential security breaches, providing a solid foundation for our analysis in this study.

3.2. RELIABILITY, VALIDITY & CONSISTENCY

Reliability: The dataset provides a detailed logging of network traffic over a consistent three-month period, suggesting a high degree of reliability in terms of data collection. The data spans from July 1 to September 30, 2006, covering network interactions across 10 local workstation IPs. The continuous recording during this period provides a robust basis for analyzing network behavior over time.

Validity: The content of the dataset is highly relevant to the research objectives, capturing essential network traffic metrics such as date, local IP, remote ASN, and flow counts. This data directly supports the investigation of network security concerns, such as the identification of anomalies and patterns indicative of security compromises. The detailed tracking of "odd" activity further validates the dataset as a valuable resource for discerning unusual network behaviors.

Consistency: The dataset maintains a uniform structure throughout its entirety, with data consistently recorded across the same four variables. This uniformity allows for reliable comparisons and analysis over time and across different network segments. The presence of both compromised and non-compromised IPs provides a balanced view, enabling comparisons and deeper insights into the factors that might influence or indicate a network compromise.

3.3. METADATA

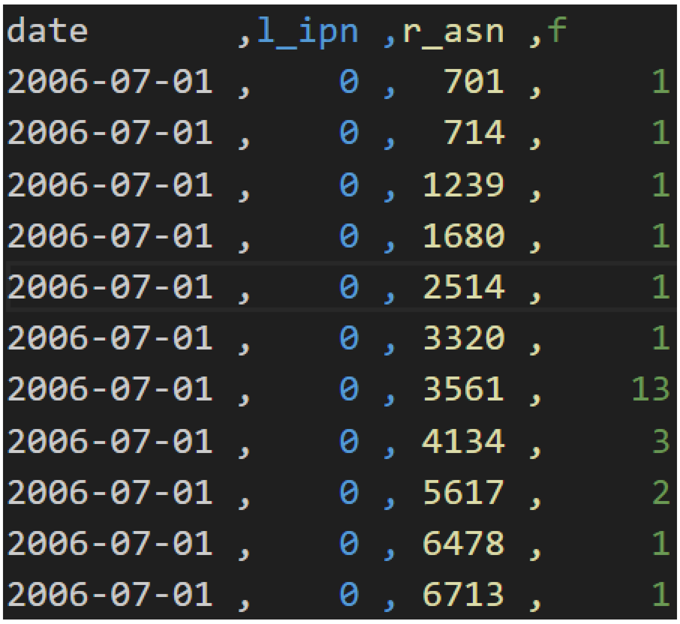
Scope: The dataset encompasses approximately 21,000 rows of data, tracking 10 local workstation IPs over a three-month period, from July 1 to September 30, 2006.

Compromises: During the observed period, half of the monitored local IPs were compromised and subsequently became part of various botnets

Content: The data is structured into four columns:

- date: The date of traffic, formatted as yyyy-mm-dd.
- l_ipn: The local IP, encoded as an integer from 0 to 9.
- r_asn: The remote Autonomous System Number, an integer identifying the remote Internet Service Provider.

- f: The flow count, representing the number of connections for the given date.



```
date       ,l_ipn ,r_asn ,f
2006-07-01 ,    0 ,  701 ,     1
2006-07-01 ,    0 ,  714 ,     1
2006-07-01 ,    0 , 1239 ,     1
2006-07-01 ,    0 , 1680 ,     1
2006-07-01 ,    0 , 2514 ,     1
2006-07-01 ,    0 , 3320 ,     1
2006-07-01 ,    0 , 3561 ,    13
2006-07-01 ,    0 , 4134 ,     3
2006-07-01 ,    0 , 5617 ,     2
2006-07-01 ,    0 , 6478 ,     1
2006-07-01 ,    0 , 6713 ,     1
```

Figure 1: Raw Data Snapshot from "Computer Network Traffic."

3.4. DATA EXPLORATION:

- The initial exploration of the dataset reveals the following:

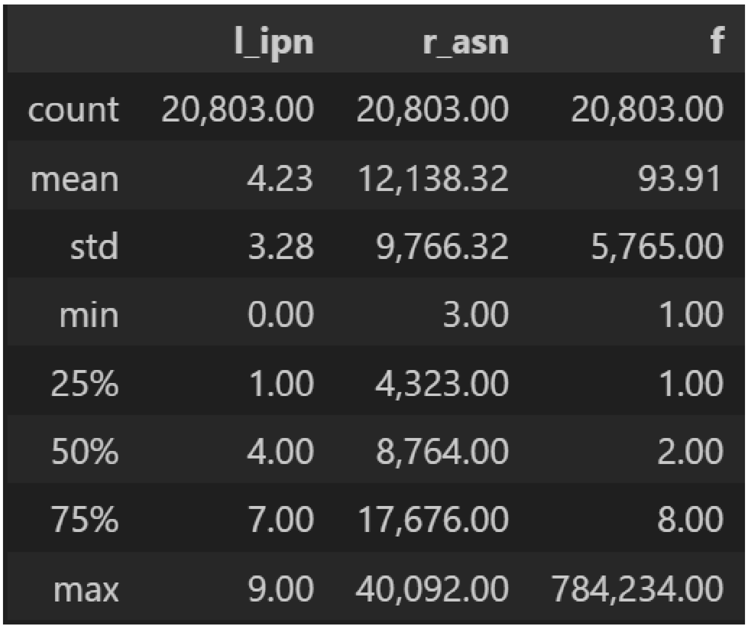
  - **Total Entries**: 20,803 records.

  - **Columns**: The dataset contains four columns: date, local IP (l_ipn), remote ASN (r_asn), and flow count (f).

  - **Local IP** (l_ipn): Ranges from 0 to 9, representing 10 local workstation IPs, with a fairly uniform distribution across these IPs.

  - **Remote ASN** (r_asn): Values range from 3 to 40,092, indicating a wide variety of remote Internet Service Providers involved in the network traffic.

  - **Flow Count** (f): The flow count varies significantly, with a minimum of 1 and a maximum of 784,234, highlighting some outliers or instances of extremely high network traffic.



	l_ipn	r_asn	f
count	20,803.00	20,803.00	20,803.00
mean	4.23	12,138.32	93.91
std	3.28	9,766.32	5,765.00
min	0.00	3.00	1.00
25%	1.00	4,323.00	1.00
50%	4.00	8,764.00	2.00
75%	7.00	17,676.00	8.00
max	9.00	40,092.00	784,234.00

Figure 2: Summary Statistics of the dataset

- **Distribution of Flow Counts**: This histogram shows the frequency of different flow counts, with the y-axis on a logarithmic scale to better represent the wide range of values. Most flow counts are low, but there are some instances with significantly higher values, indicating potential outliers or specific events causing spikes in network traffic.

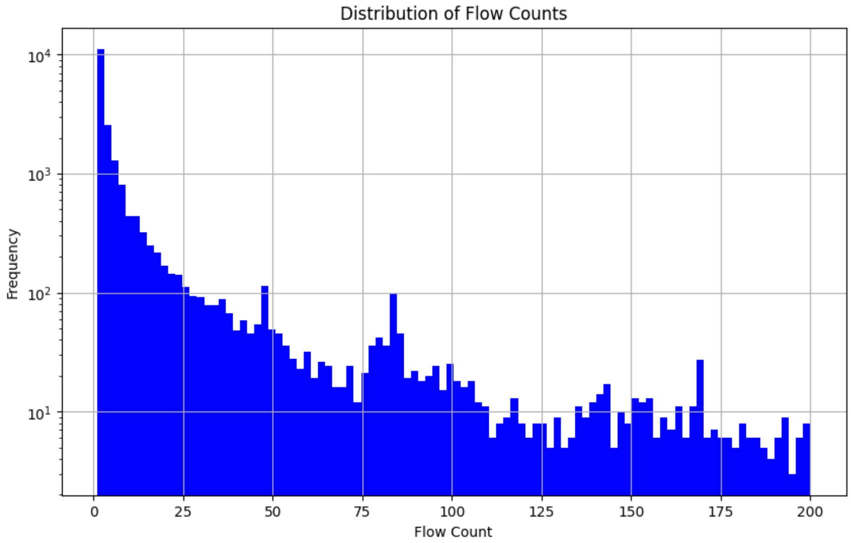


Figure 3: Distribution of flow counts.

- Network Traffic Over Time: The line graph represents the total flow count overtime, from July to September 2006. There are noticeable fluctuations in the network traffic, which could correspond to specific network events or attacks.

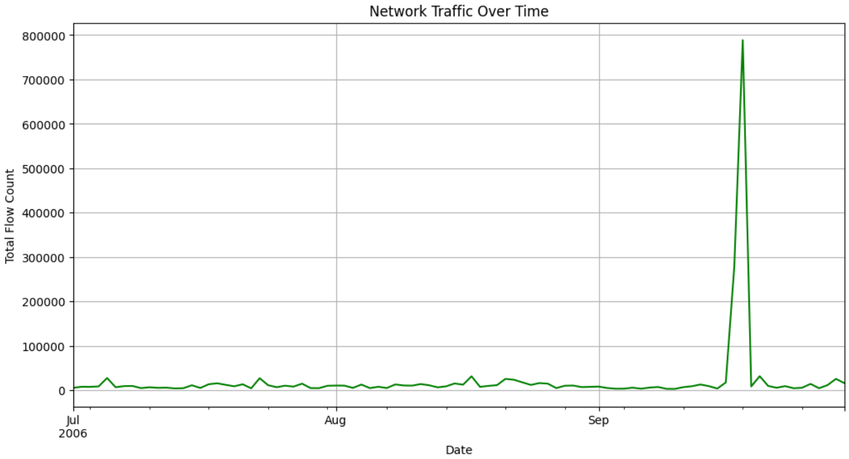


Figure 4: Network Traffic over Time

4. DYNAMIC STOCHASTIC PROCESSES & SIMULATIONS

4.1. FORMULATION

4.1.1. OBJECTIVE

The aim is to develop a stochastic model that classifies network traffic flow as 'Low', 'Medium', or 'High' to predict daily connections and identify potential security breaches.

4.1.2. STATE SPACE:

Let S = {Low, Medium, High} be the state space representing traffic flow levels.

4.1.3. STATE DIAGRAM:

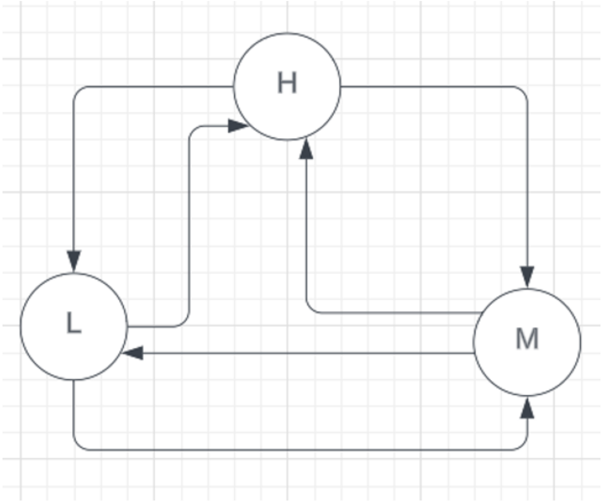


Figure 5: Transition Diagram

4.1.4. TRANSITION MATRIX:

We define the transition matrix P as follow:

$$P = \begin{bmatrix} p_{LL} & p_{LM} & p_{LH} \\ p_{ML} & p_{MM} & p_{MH} \\ p_{HL} & p_{HM} & p_{HH} \end{bmatrix}$$

Figure 6: Transition Matrix

4.1.5. DECISION CRITERION BASED ON STATIONAR DISTRIBUTION

The stationary distribution $\pi = \{ \pi_L, \pi_M, \pi_H \}$ gives the long-term probabilities of the system being in 'Low', 'Medium', or 'High' states, respectively.

The system is deemed compromised if:

- π_H exceeds a predefined security threshold, indicating a high probability of persistently high traffic, which could signify an ongoing attack..

4.2. SIMULATION: REALIZATION OF USED PROCESSES

4.2.1. METHOD 1: SIMULATION OVER TIME

Objective: To simulate the Markov chain representing network traffic flow with 'Low', 'Medium', and 'High' states over a certain period.

Method: We will create a function that simulates the state transitions over a specified number of days, starting from an initial state.

Hypothetical Transition Matrix:

$$\begin{matrix} \mathbf{0.7} & \mathbf{0.2} & \mathbf{0.1} \\ \mathbf{0.3} & \mathbf{0.4} & \mathbf{0.3} \\ \mathbf{0.2} & \mathbf{0.3} & \mathbf{0.5} \end{matrix}$$

Initial State: We will assume the system starts in the 'Low' state.

Results:

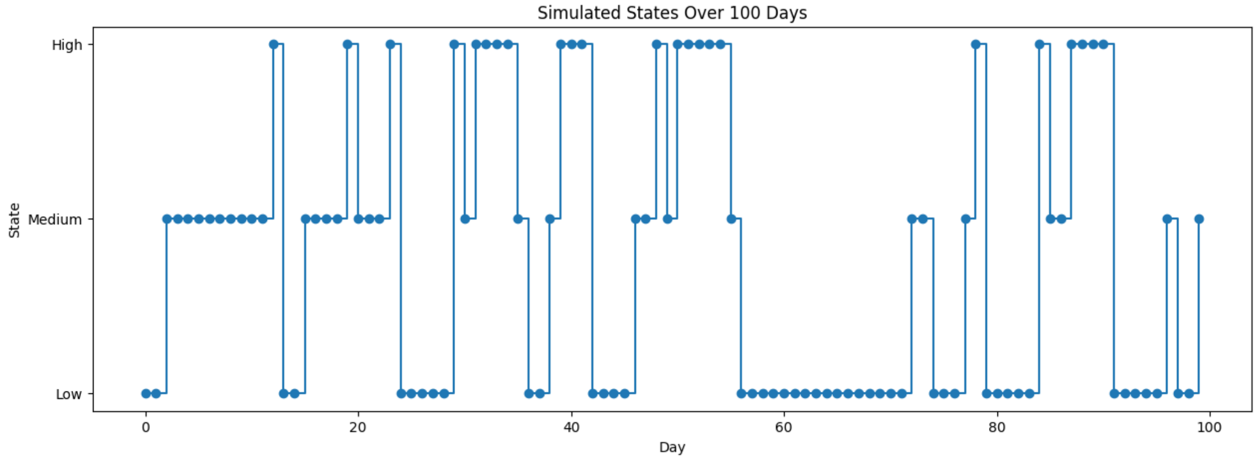


Figure 7: Simulated States Over 100 Days

The stepping pattern demonstrates the stochastic nature of the system, with changes in state occurring at random intervals dictated by the underlying transition probabilities. The plot shows frequent oscillation between all three states, with no single state showing dominance over the others in terms of the length of time spent.

4.2.2. METHOD 2: MONTE CARLO SIMULATION:

Objective: To estimate the long-term behavior of the network traffic flow states using Monte Carlo simulation

Method: Run a large number of Markov chain simulations and record the frequency of each state at the end of the simulations to estimate the stationary distribution.

Results: The estimated stationary distribution after 10,000 simulations is as follows:

- Low state: 0.4558
- Medium state: 0.2738
- High state: 0.2704

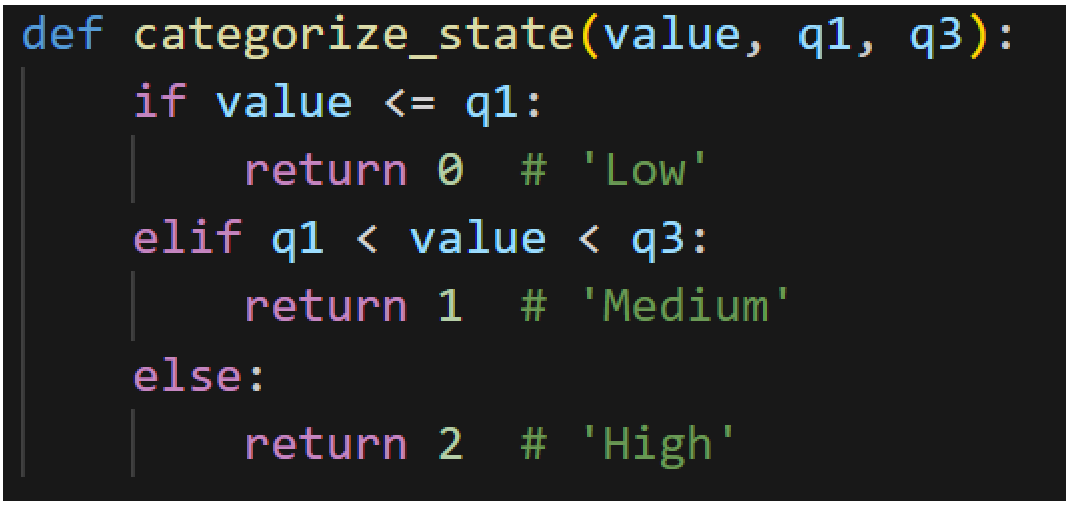
Analysis:

- The 'Low' state is the most probable long-term state, with a likelihood of approximately 45.58%. This suggests that over a significant period, the network is more likely to experience low traffic.
- The 'Medium' and 'High' states have a lower probability, 27.38% and 27.04% respectively. However, the relatively high percentage of the 'High' state indicates a non-negligible risk of potential security breaches, as high traffic flow might be associated with system compromises.

4.3. MODELING USING REAL DATA

We will categorize the flow counts (f) into discrete states (e.g., "Low", "Medium", "High") based on percentile thresholds. This categorization transforms the continuous flow data into a format suitable for a Markov Chain analysis, where each state represents a level of network traffic activity.

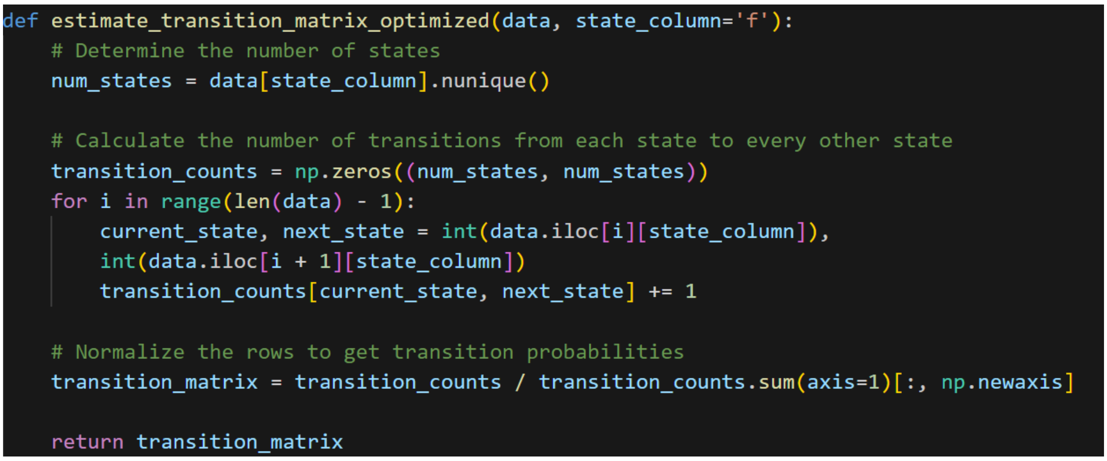
- 'Low' corresponds to the values less than or equal to Q1 (1st quartile),
- 'Medium' corresponds to the values greater than Q1 and less than Q3 (3rd quartile),
- 'High' corresponds to the values greater than or equal to Q3



```python
def categorize_state(value, q1, q3):
    if value <= q1:
        return 0  # 'Low'
    elif q1 < value < q3:
        return 1  # 'Medium'
    else:
        return 2  # 'High'
```

Figure 8: Categorizing States

Next, we estimate the probabilities of transitioning from one state to another:



```python
def estimate_transition_matrix_optimized(data, state_column='f'):
    # Determine the number of states
    num_states = data[state_column].nunique()

    # Calculate the number of transitions from each state to every other state
    transition_counts = np.zeros((num_states, num_states))
    for i in range(len(data) - 1):
        current_state, next_state = int(data.iloc[i][state_column]),
        int(data.iloc[i + 1][state_column])
        transition_counts[current_state, next_state] += 1

    # Normalize the rows to get transition probabilities
    transition_matrix = transition_counts / transition_counts.sum(axis=1)[:, np.newaxis]

    return transition_matrix
```

Figure 9: Function to generate transition matrix from the dataset

Finally, we get the following transition matrix:

$$\begin{matrix} \mathbf{46.1} & \mathbf{36.12} & \mathbf{17.78} \\ \mathbf{30.14} & \mathbf{46.16} & \mathbf{23.7} \\ \mathbf{25.1} & \mathbf{41.3} & \mathbf{33.6} \end{matrix}$$

4.4. ESTIMATION: MAXIMUM LIKELIHOOD

Maximum likelihood estimation (MLE) will involve defining a likelihood function based on the transition probabilities and the observed data, then finding the parameter values that maximize this likelihood.

For a Markov chain, the likelihood L is of observing a sequence of states given the transition matrix P:

$$L(P) = \prod_{t=1}^{n-1} p_{x_t x_{t+1}}$$

We will need to iterate over our dataset and compute the product of the observed transitions. In practice, it is more common to work with the log-likelihood because it transforms the product into a sum, making it easier to work with:

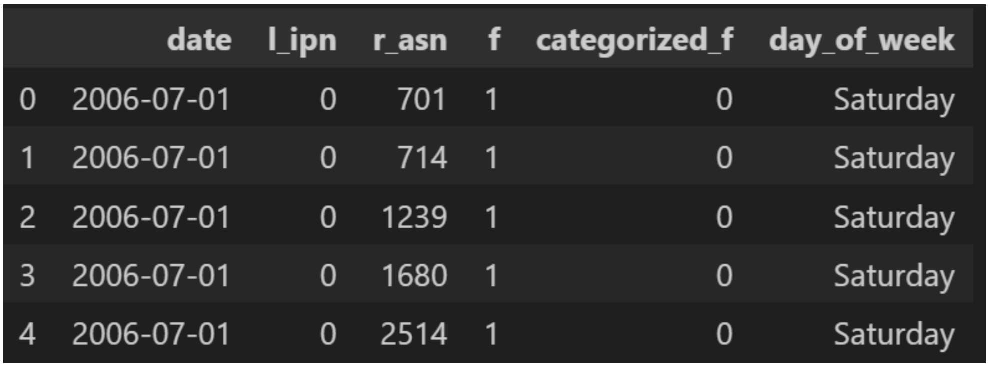
$$\log(\,L\,(P)) = \sum_{t=1}^{n-1} \log\left(p_{x_t x_{t+1}}\right)$$

Model Fit: A log-likelihood of -22363.70 suggests that while the model captures the general pattern of transitions, there may be room for improvement. Adjustments to the model or using a more refined method for categorizing states could potentially yield a better fit. Also, the log-likelihood is influenced by influenced by the number of data points simply due to the accumulation of probabilities.

4.4.1. Further Exploration

We can convert the date column into more granular time components such as day of the week or part of the day if hourly data is available. This could reveal patterns like higher traffic on specific days or times, which could be crucial for understanding traffic flow dynamics.

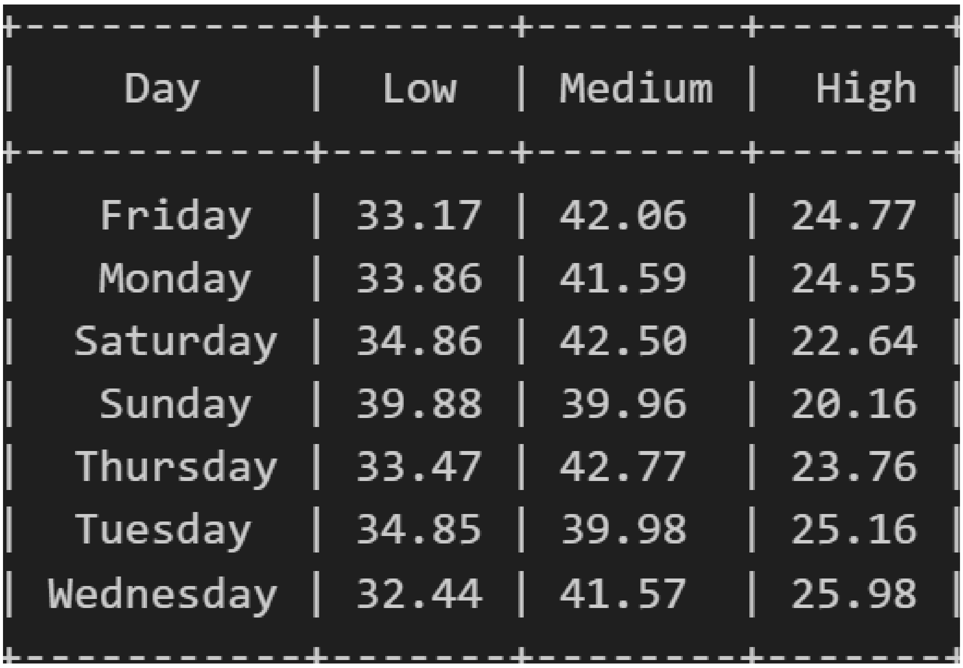
Let's start with a temporal analysis by extracting the day of the week from the date column, which could potentially uncover weekly traffic patterns. We will then see if there's a variation in state transitions based on the day of the week.



	date	l_ipn	r_asn	f	categorized_f	day_of_week
0	2006-07-01	0	701	1	0	Saturday
1	2006-07-01	0	714	1	0	Saturday
2	2006-07-01	0	1239	1	0	Saturday
3	2006-07-01	0	1680	1	0	Saturday
4	2006-07-01	0	2514	1	0	Saturday

Figure 10: Adding the day of week into the dataset

Next, we will be calculating the frequency of each traffic state ('Low', 'Medium', 'High') for each day of the week. This will provide an initial understanding of how traffic states might be influenced by the day of the week.



```
+-----------+-------+--------+-------+
|    Day    |  Low  | Medium |  High |
+-----------+-------+--------+-------+
|   Friday  | 33.17 | 42.06  | 24.77 |
|   Monday  | 33.86 | 41.59  | 24.55 |
|  Saturday | 34.86 | 42.50  | 22.64 |
|   Sunday  | 39.88 | 39.96  | 20.16 |
|  Thursday | 33.47 | 42.77  | 23.76 |
|  Tuesday  | 34.85 | 39.98  | 25.16 |
| Wednesday | 32.44 | 41.57  | 25.98 |
+-----------+-------+--------+-------+
```

Figure 11: Frequency of each traffic for each day of the week.

Chi-Squared Test Results: The chi-squared test was performed to determine if there is a statistically significant difference in the frequency of traffic states ('Low', 'Medium', 'High') across different days of the week. Here are the results:

- Chi-Squared Statistic: 60.54
- P-Value: 1.34 e-07

Interpretation:

- The chi-squared statistic is relatively high, and the p-value is very low (far below the common threshold of 0.05), indicating that we can reject the null hypothesis.

- This suggests that the differences in traffic state distributions across different days of the **week are statistically significant.**

Conclusions: There is a statistically significant difference in how traffic states are distributed across different days, which implies that day-specific patterns exist in network traffic.

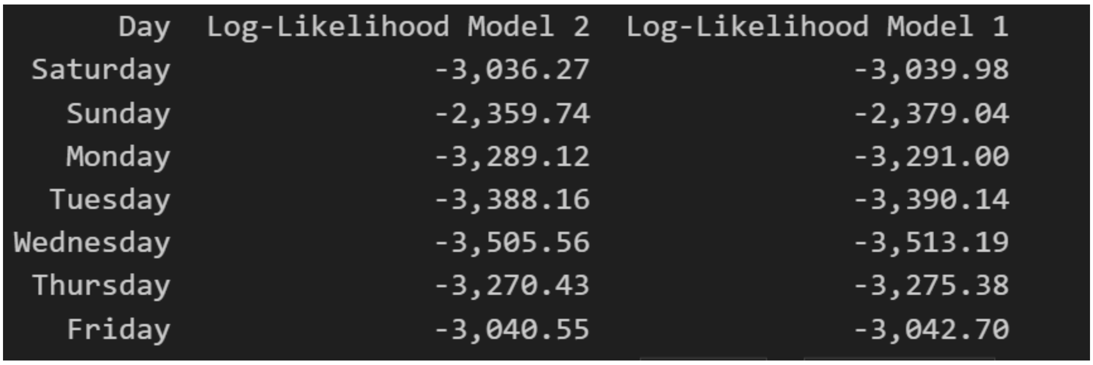
REFINED MODEL:

Statistical analysis (chi-squared test) revealed significant differences in traffic states across different days of the week, suggesting the inclusion of day-specific patterns could enhance model accuracy.

Transition probabilities were recalculated to incorporate day of the week, allowing the model to capture daily variations more effectively.

4.4.2. COMPARING BOTH MODELS:

The comparison between Model 1 and Model 2, as shown in the table, indicates a marginal improvement in log-likelihood values across all days of the week for Model 2, which incorporates day-specific transition matrices. This improvement suggests that Model 2 is a slightly better fit for the data, reflecting the importance of day-of-week effects in predicting network traffic states. Notably, the increase in log-likelihood values is consistent, though modest, reinforcing the hypothesis that network traffic patterns have day-dependent characteristics.



```
      Day  Log-Likelihood Model 2  Log-Likelihood Model 1
 Saturday               -3,036.27               -3,039.98
   Sunday               -2,359.74               -2,379.04
   Monday               -3,289.12               -3,291.00
  Tuesday               -3,388.16               -3,390.14
Wednesday               -3,505.56               -3,513.19
 Thursday               -3,270.43               -3,275.38
   Friday               -3,040.55               -3,042.70
```

Figure 12: Comparison between model 1 and 2.

4.5. INTERPRETATION

From the estimated transition probabilities and the analysis incorporating day of the week, we observed significant variations in traffic states, which have direct implications for cybersecurity:

High Traffic State Probability:

- Days with statistically higher probabilities of transitioning to or remaining in a 'High' state should be of particular concern as these might indicate not just high traffic but potential security vulnerabilities or ongoing attacks.

- For instance, if the probability of staying in or transitioning to a 'High' state on a Wednesday is significantly greater than on other days, this suggests a pattern that could be related to specific weekly activities or system vulnerabilities exploited at midweek.

Stationary Distribution Insights:

- The long-term probabilities derived from the stationary distribution indicate how often, on average, the network is expected to be in each state. A high long-term probability for the 'High' state across several days suggests an enduring vulnerability or insufficient capacity to handle peak loads, both of which could compromise system security.

Day-Specific Patterns:

- Days showing a consistent shift towards higher traffic states, based on the chi-squared test results and the refined transition matrices, should prompt targeted security checks and preparedness. For instance, increased monitoring and preemptive threat detection measures could be scheduled in anticipation of these high-risk periods.
- Predictive Security Measures: Use the model’s output to implement predictive security measures. On days predicted to have a high likelihood of 'High' traffic states, enhance surveillance and intrusion detection sensitivity.
- Resource Allocation: Allocate more bandwidth and cybersecurity resources on days with predicted high traffic to prevent system overloads and potential breaches.
- Incident Response Planning: Develop rapid response plans for days identified as high risk, ensuring that response teams are on alert to mitigate any potential security incidents quickly.

4.6. IMPROVEMENT STRATEGIES

- Incorporate External Data: Integrate external factors such as known cyber-attack patterns, maintenance schedules, or significant external events (e.g., product launches, major announcements) that might affect traffic patterns.
- Utilize machine learning techniques, such as random forests or neural networks, to predict high traffic states more dynamically, incorporating a broader range of inputs.
- Conduct red team exercises based on the model’s predictions to test system defenses against simulated attacks on high-risk days.

- Use of dynamic data while generating the transition matric by using the daily logs.

5. FURTHER DEVELOPMENTS: RESEARCH & SELF-LEARNING

As we look to the future, continuous refinement of the model will be critical. Research will delve into more sophisticated machine learning algorithms, such as deep learning for pattern recognition in traffic flow, which can adapt to new threats more dynamically. Incorporating real-time data analytics will pave the way for a more proactive defense mechanism. Self-learning systems using reinforcement learning could revolutionize the model's ability to predict and respond to threats autonomously, learning from past experiences and adjusting their prediction strategies without human intervention.

6. CONCLUSION

In conclusion, our stochastic model's journey—from its initial conception to the integration of day-specific transition matrices—highlights the essence of iterative development in the realm of data-driven predictive analytics. The incremental improvements observed by refining the model using maximum likelihood estimations and moment methods underscore the value of a rigorous, data-centric approach. As we continue to hone our model, its role in enhancing cybersecurity measures remains indispensable, promising a more resilient network environment capable of withstanding the evolving landscape of digital threats.

REFERENCES

- Silva, L. D., Lollini, P., Mongelli, D., Bondavalli, A., & Mandò, G. (2021). A stochastic modeling approach for traffic analysis of a tramway system with virtual tags and local positioning. Journal of the Brazilian Computer Society, 27(2). https://doi.org/10.1186/s13173-021-00105-x
- *Computer network traffic*. (2017, August 18). Kaggle. https://www.kaggle.com/datasets/crawford/computer-network-traffic
- Jabari, S. E., & Liu, H. (2012). A stochastic model of traffic flow: Theoretical foundations. *Transportation Research Part B: Methodological*, *46*(1), 156–174. https://doi.org/10.1016/j.trb.2011.09.006
- Anonymous. (2009, January 29). Maximum likelihood estimation for Markov chains. Lecture notes to accompany lecture 6 in course 36-462, Spring 2009. https://www.stat.cmu.edu/~cshalizi/462/lectures/06/markov-mle.pdf