analytics

MDPI

Article

Cyberpsychology: A Longitudinal Analysis of Cyber Adversarial Tactics and Techniques

Marshall S. Rich

Cyberpsychology, Capitol Technology University, Laurel, MD 20708, USA; mrich@captechu.edu or marshall.rich@richoncyber.com; Tel.: +1-478-747-3374

Abstract: The rapid proliferation of cyberthreats necessitates a robust understanding of their evolution and associated tactics, as found in this study. A longitudinal analysis of these threats was conducted, utilizing a six-year data set obtained from a deception network, which emphasized its significance in the study's primary aim: the exhaustive exploration of the tactics and strategies utilized by cybercriminals and how these tactics and techniques evolved in sophistication and target specificity over time. Different cyberattack instances were dissected and interpreted, with the patterns behind target selection shown. The focus was on unveiling patterns behind target selection and highlighting recurring techniques and emerging trends. The study's methodological design incorporated data preprocessing, exploratory data analysis, clustering and anomaly detection, temporal analysis, and cross-referencing. The validation process underscored the reliability and robustness of the findings, providing evidence of increasingly sophisticated, targeted cyberattacks. The work identified three distinct network traffic behavior clusters and temporal attack patterns. A validated scoring mechanism provided a benchmark for network anomalies, applicable for predictive analysis and facilitating comparative study of network behaviors. This benchmarking aids organizations in proactively identifying and responding to potential threats. The study significantly contributed to the cybersecurity discourse, offering insights that could guide the development of more effective defense strategies. The need for further investigation into the nature of detected anomalies was acknowledged, advocating for continuous research and proactive defense strategies in the face of the constantly evolving landscape of cyberthreats.

Keywords: honeypots; intrusion detection; threat hunting; threat intelligence; cyberdefense; cybersecurity; cyberattacks; cyberthreats; network security

check for updates

Citation: Rich, M.S. Cyberpsychology: A Longitudinal Analysis of Cyber Adversarial Tactics and Techniques. *Analytics* **2023**, *2*, 618–655. https://doi.org/10.3390/analytics2030035

Academic Editor: Alessandro Provetti

Received: 30 June 2023
Revised: 31 July 2023
Accepted: 9 August 2023
Published: 11 August 2023

Copyright: © 2023 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

1.1. The Problem

The primary purpose of this study is to perform a longitudinal analysis with the goal of performing an exhaustive exploration of the tactics and strategies utilized by cybercriminals and how these tactics and techniques evolved in sophistication and target specificity over time.

The cybersecurity ecosystem continues to evolve rapidly, challenging practitioners and researchers to keep pace with these changes [1,2]. Cyberattack techniques are diversifying, and the threat landscape is expanding, with new types of network-based cyberattacks being discovered [3]. Researchers like Weathersby have explored the relative threats of these various cyberattack forms and understand hackers' motivation, attribution, and anonymity [3]. Similarly, Myneni emphasizes bridging the gap between security and intelligence to defeat attackers more effectively [4].

Effectively countering cyberthreats calls for a comprehensive approach that encapsulates the perspectives of all stakeholders, including defenders, attackers, and bystanders [3,4]. This holistic approach not only provides an all-encompassing analysis of the cyberthreat

landscape but also aids in the formulation of robust defensive strategies [5–7]. Technological evolution, while offering numerous advancements and opportunities, also presents a myriad of pathways for malicious exploits and threat actors [4,8–10].

Examining cyberthreat trends and patterns necessitates thoroughly comprehending deception networks, honeypots, and decoy systems [6,11,12]. These systems are instrumental in intelligence gathering and serve as deterrents to cyberattacks, and have shown effectiveness in monitoring network traffic and identifying potential threats [6,11–13]. Their capacity for intelligence gathering against targeted systems and networks underlines their crucial role in cyberdefense [6,11,14].

Open-Source Intelligence (OSINT) is another significant tool for identifying and forecasting cyberthreats [8,15,16]. Combining OSINT and deception intelligence information provides a comprehensive and multifaceted approach to cybersecurity. This approach not only leverages the proactive capabilities of deception networks, honeypots, and decoy systems to detect and deter cyberthreats but also capitalizes on the extensive reach and forecasting abilities of OSINT [8,15,16]. Together, these tools allow for a broader and more informed understanding of cybercriminal strategies and potential threats, thus enhancing the effectiveness of proactive defense mechanisms, identifying threats, and developing robust and adaptable cybersecurity measures [6,11–14,16].

This study is centered around a longitudinal analysis of cyberthreats harvested from a deception network, meticulously tracking the evolution of threat actors' tactics and strategies over a designated period. This time-bound examination provides a narrative on the progress of specific threats. It illustrates the broader shifts in the cybercrime landscape, indicating how threats have become increasingly complex and adaptive [4,11,17].

The study augments the data collection and enhances the depth of its analysis by incorporating OSINT derived from public threat intelligence data feeds. OSINT allows the fusing of multiple data layers, giving the analysis a richer, more expansive perspective. It validates the findings from the deception network and helps identify correlations and intersections between separate datasets, leading to a more comprehensive understanding of cyberthreats and their trajectory [6,16,18].

The paper strives to shed light on the growing sophistication of these cyberthreats. It delves into the intricate details of advanced attack methods, revealing how threat actors continuously innovate to bypass security measures. Doing so underscores the urgent need for adaptable, multifaceted cybersecurity strategies to keep pace with these evolving threats [3,19,20].

The targeted specificity of these cyberthreats is also explored in this study. Through a careful dissection of distinct instances of attacks, patterns behind target selection can be understood, irrespective of whether they pertain to sectors, geographies, or exploits. This analysis could enable organizations to anticipate forthcoming threats and fortify their defenses accordingly [19,21,22].

Lastly, an investigation is undertaken into the categories of threats that have observed a significant surge. This mixed-method analysis, encompassing qualitative and quantitative approaches, provides valuable insights into the current and emerging areas of intense cybercriminal activity. Whether it is an escalating frequency of ransomware attacks, an uptick in phishing campaigns, or the proliferation of Advanced Persistent Threats (APTs), gaining comprehension of these trends can assist stakeholders in prioritizing their cybersecurity initiatives and responding more effectively to imminent threats [3,13,23].

Collectively, these varied investigations aim to enhance the universal understanding of the multifaceted and dynamic landscape of cyberthreats, yielding crucial insights for the academic domain and the practical sphere of cybersecurity [5,15,19,20].

Given the complex nature of cyberthreats, innovative methodologies are required. The incorporation of artificial intelligence (AI) and machine learning (ML) techniques in threat detection strategies has shown promise [24–28]. Furthermore, cyberdeception methods have shown the potential to address the imbalance between attackers and defenders [6,12].

The continuously evolving landscape of cyberthreats demands ongoing research and understanding [2,3,11,19]. This study aims to enrich the existing literature by providing deep insights into cybersecurity issues through a detailed analytical analysis of changing cyberattack patterns and trends [2,11,19,29]. By enhancing our understanding of dynamic cyberattack strategies, this exploration aims to offer valuable insights that could aid in developing adaptable and resilient cybersecurity measures [4,7,17].

1.2. Objective and Scope

The main objective of this research article is to perform a rigorous and systematic examination of the tactics and strategies employed by cybercriminals. This investigation will be informed by a six-year data set acquired from a deception network. A comprehensive longitudinal analysis is the primary focus of this study, aiming to reveal pivotal trends, patterns, and the evolutionary nature of cyberattack methods [6,8,14,30]. These findings seek to stimulate an in-depth understanding of the persistently evolving tactics of cybercriminals, offering insights with the potential to bolster cybersecurity strategies [5,15,16,19] significantly.

Regarding the scope, this research will probe into the activities of cybercriminals as mirrored in the dataset, which covers a wide range of cyberattacks, including but not limited to intrusion attempts and APTs [3,10,13,23]. Integrating OSINT from public malicious threat feeds will augment the richness and reach of the analysis, providing a broader perspective on cyberthreat trends [8,15,16].

Overall, this research is designed to deliver practical insights that could facilitate the design and implementation of robust cybersecurity measures. It also seeks to contribute significantly to the existing academic discourse on cybersecurity and analytics [5,15,16,19]. The goal is to enhance our collective understanding of the dynamic tactics utilized by cybercriminals and provide critical insights that could inform the development of resilient and adaptable cybersecurity strategies. Furthermore, the research emphasizes threats' escalating sophistication, targeted specificity, and categories that have witnessed a notable increase [3,19,23].

1.3. Research Question and Hypothesis

In line with the objective and scope of this research, the main question propounded for investigation is:

- Research Question (RQ1): What are the key trends and patterns in cyberattacks over the analyzed period, and how have these tactics and techniques evolved in sophistication and target specificity over time?

This question is directed towards examining the evolving landscape of cyberthreats, focusing on developing their sophistication and precision in target selection. This probe is expected to expose the trajectories of these cyberattacks, subsequently informing the creation of future cybersecurity strategies [5,6,14,16,19].

From the research question, the subsequent hypothesis is advanced for evaluation in this study:

Hypothesis (H1): *Cyberattacks are becoming increasingly sophisticated and targeted over time, with certain types of attacks showing a marked increase.*

This hypothesis is rooted in studies illustrating the evolution of cyberthreats [8,14,23,30], emphasizing the unceasing necessity for enhancing cybersecurity measures. The presupposition is that the tactics and techniques of cybercriminals have escalated in their sophistication, particularly in terms of their ability to breach targeted systems and networks [6,11,14]. To validate this hypothesis, an in-depth analysis of a six-year dataset from a deception network will be performed, monitoring the transformation of cyberattacks, their escalating intricacy, and the evolving strategies adopted by cybercriminals.

If the findings corroborate this hypothesis, it will accentuate the importance of continual progress and adaptability in cybersecurity measures. More specifically, it would highlight the significance of deception networks, decoy systems, and honeypots in collecting intelligence against targeted systems and networks for defense [6,11,14,16].

1.4. Significance of the Research

The importance of this research unfolds in three ways. First, by offering a detailed longitudinal analysis of the tactics and methods employed by cybercriminals over six years, this study aims to enrich the existing body of cybersecurity and analytical literature with valuable insights [5,15,19,20]. With the perpetual evolution of cyberthreats, continuous research in this domain is crucial to keep abreast of these changes, ensuring that defensive strategies are updated and effective. This research's emphasis on revealing key trends, patterns, and alterations in cyberattack strategies is geared toward addressing an identified gap in the existing academic literature [6,8,21,30].

Second, the practical relevance of this research is noteworthy. The understanding obtained from scrutinizing a six-year dataset from a deception network could guide the formation of adaptive and forward-looking cybersecurity strategies. In a time where digital connectivity is increasing, and new attack vectors are incessantly emerging, especially with the advent of Internet of Things (IoT) networks [16,30,31], the results of this study could influence security protocols across a multitude of organizations and could enhance their ability to detect, counter, and preemptively prevent cyberthreats.

Lastly, the focus on the efficacy of deception networks, decoy systems, and honeypots in collecting defensive intelligence against targeted systems and networks [5,11,14] further underscores the significance of this research. Considering the expected outcomes, this research could encourage broader and more efficient employment of these tools, enhancing collective cybersecurity defense strategies at the organizational, national, and international levels.

In conclusion, with its theoretical contributions and practical implications, this research holds substantial value for academia and industry, fostering an understanding of cyberthreats and promoting effective mitigation strategies.

2. Materials and Methods

2.1. Methodology

The study's methodological approach was designed to address RQ1, focused on unearthing vital trends and patterns in cyberattacks over six years and analyzing the advancement in sophistication and target specificity of these techniques. This exploration was underlined by a strong focus on longitudinal analysis, which lends itself to a comprehensive examination of the tactics and strategies employed by cybercriminals [15,20].

The study followed Hypothesis (H1), proposing a rising sophistication and specificity in cyberattacks, with certain attack types exhibiting a discernible increase. The investigative methodology was built on an analytical process featuring stages such as data preprocessing, exploratory data analysis, clustering and anomaly detection, temporal analysis, cross-referencing with known threat intelligence, visualization and reporting, and interpretation and implications [22,32]. This method is anticipated to generate crucial insights to strengthen cybersecurity strategies in a world marked by growing digital interconnectivity [16,30,31].

2.2. Data Collection

A sizable dataset was acquired from a deception network log that spanned six years; this long-term collection period generated an extensive quantity of records, permitting an exhaustive analysis of the tactics and techniques employed by cybercriminals. The sheer magnitude of this dataset underscored its value for this research, proving it to be an invaluable resource for this investigation [14,30]. No changes to the deception network configuration were introduced during these six years. The Internet attack-surface Internet

Protocol (IP) address, routes, and firewall configurations remained unchanged. In October 2022, the project concluded, and the deception network was removed from the Internet.

The deception network or honeypot system leveraged in this study was designed to emulate various services and systems frequently targeted by cyberattackers. This strategy ensured a comprehensive record of attacks, capturing a diverse array of tactics and techniques utilized by cybercriminals. Operational from October 2016 to September 2022, the honeypot was engaged in constantly collecting data related to attempted and unsuccessful breaches.

This dataset, a product of a deception network, emphasized its significance in the study's primary aim: the exhaustive exploration of the tactics and strategies utilized by cybercriminals [18,20]. The depth of the data collated offered a sturdy platform for a longitudinal analysis to reveal critical trends, patterns, and evolutions in cyberattack methods [18], thereby supporting both RQ1 and Hypothesis (H1).

The utilization of honeypot-derived data presents a unique vantage point into cybercriminals' actual tactics and techniques, offering crucial insights into the evolution of cyberthreats. Consequently, the data collection methodology employed by this study establishes a robust basis for examining trends and patterns in cyberattacks [11,22,30].

Original Dataset Log Format and Description

The deception network or honeypot system logged each connection attempt, whereas each row of the log file corresponded to a single event. Each event has multiple pieces of information associated with it. Based on the structure and content of the logs, here is what each column represents (this log file will be referred to as "file 1"):

- Event Timestamp: The date and time of each event were captured in this column, playing a vital role in event chronology and correlation.
- Communication Protocol Type: The protocol utilized for network communication was indicated here. "tcp" was indicative of the Transmission Control Protocol (TCP).
- Protocol Identifier: Represents the protocol number designated by the Internet Assigned Numbers Authority (IANA) for protocol identification. "6" was associated with TCP.
- Nature of Event: The type of event was indicated here. "S" represented a SYN (synchronize) packet utilized to establish a TCP connection, while "E" represented the termination or an error within the connection.
- Source IP Address: This represents the system's IP address from which the network traffic originated.
- Source Port Number: The specific port on the originating system dispatching the network traffic was documented here.
- Destination IP Address: The IP address of the target system to which the network traffic was directed is indicated here.
- Destination Port Number: This was the port on the destination system receiving the network traffic.
- Packet Details: Represented packet-related details, such as flags or length. Within this log, the constant "0 0" signify the initiation or termination of a connection.

2.3. Data Preprocessing

The initiation of data preprocessing involved cleansing and conditioning the log data extracted from the honeypot. Processes included eliminating irrelevant or redundant entries, addressing missing values, and formatting the data to make it suitable for subsequent analysis. Key data points such as timestamp, source IP address, destination port, and destination service were extracted during this stage.

Furthermore, data normalization was performed, a necessary step for maintaining consistency in data scales, facilitating more effective analysis. Such normalization is crucial when dealing with datasets encompassing a variety of variables and measurements [14,21].

A crucial preprocessing aspect was feature extraction, wherein relevant attributes were identified and extracted from the dataset to support the intended investigation. For instance, timestamp data were disaggregated to yield additional insights such as the date, day of the week, and month when the attack was initiated.

Data transformation formed a vital part of the preprocessing phase as well. This step facilitated the conversion of the data into a format suitable for analysis, transforming categorical data into numerical representations where required and encoding specific data to ease the analysis process.

Post-preprocessing, the cleansed dataset was primed for a comprehensive examination. This examination involved probing for patterns, trends, correlations, and behavioral insights within the data using various analytical instruments and methodologies [14,24,33].

Given the potentially intensive computational requirements of processing the extensive honeypot log, significant computational resources and expertise were deployed during this research phase. This measure reflected the study's commitment to maintaining rigorous data management standards, even with large datasets [1,28].

2.3.1. Data Anonymization and Transformation

For the assurance of data integrity and relevance of "file 1" in the subsequent analysis, additional preprocessing was executed on the extracted log data. The procedure encompassed (the output of this log file will be referred to as "file 2"):

- Normalization of timestamp formats.
- Extraction and conversion of timestamp information into a numerical format.
- Transformation of the source IP address into a numerical format.
- Removal of irrelevant entries (Communication Protocol Type, Protocol Identifier, Nature of Event, Source Port Number, and Packet Details).
- Conversion of any remaining categorical data into numerical equivalents.
- Omission of the target system's IP address (Destination IP Address) where the network traffic was directed. This field was removed as only one target system was evaluated in this study.

2.3.2. GeoIP and Destination Service Enrichment

Following the "Data Anonymization and Transformation" (Section 2.3.1) and utilizing "file 2", the log underwent further processing to incorporate additional context about the geographical origin of the source IP, the associated organization, and the identification of the destination service. This enriched data will later be employed with "Malicious Threat Intelligence Feeds", as it provides deeper insight into potential actors behind the activity and their intentions.

The columns within this enriched log file can be delineated as follows (the output of this log file will be referred to as "file 3"):

- Source_IP: Denotes the IP address of the originating system for the network traffic.
- Country Name: Indicates the name of the country where the source IP is situated.
- Country_ISO_Code: Represents the ISO 3166-1 alpha-2 code corresponding to the country of the source IP.
- Country_Numeric_Code: Specifies the ISO 3166-1 numeric code related to the country of the source IP.
- Continent_Code: Provides the two-letter code for the continent housing the source IP.
- Continent_Numeric_Code: Represents the numeric code assigned to the continent of the source IP.
- Autonomous System Number (ASN): Refers to the unique ASN linked with the source IP address. An ASN functions as a unique identifier of a network on the internet.
- ASN_Organization: This is the name of the organization or entity owning the ASN.
- Destination_IP: Denotes the IP address of the system receiving the network traffic.
- Destination_Port: Specifies the port number on the receiving system to which the network traffic is directed.

- Service_Name: Represents the name of the destination service used in the connection (examples include: DNS, SSH, Telnet, HTTP, and HTTPS).

2.3.3. Enrichment through "Malicious Threat Intelligence Feeds"

Definition and Explanation

This study utilized diverse "Malicious Threat Intelligence Feeds" as a primary data source. These feeds, which represent a continuous stream of information, provide crucial insights about potential or current threats to information systems and networks. The data gathered from internal and external sources paints a comprehensive picture of current cybersecurity indicators of compromise (IOC), highlighting threats such as new malware variants, phishing campaigns, and active exploits.

An example of the feeds incorporated in the study, as shown in Table 1, demonstrated various data collection methods. For instance, "ip-firehol-anonymous" is a feed that collects data about potentially threatening anonymous individual IP addresses of hosts on the Internet. At the same time, "subnet-misp-bro" is a feed that gathers threat intelligence data differently by using an IP subnet or by the network, and "actor_indicators_apt28-ipv4" is a feed that has specific IOCs from known specific threat actor or group "APT 28".

Table 1. Sample Data from Threat Intelligence Feeds.

Ip-Firehol-Anonymous	Ip-Dan-Torlist-Exit-Ip	Subnet-Misp-Bro	Subnet-Uceprotect-Dnsbl-3	Actor_Indicators_Apt28-Ipv4
100.0.0.1	102.130.113.9	101.134.0.0/15	100.1.1.0/24	101.99.83.131
100.0.0.63	102.130.127.117	101.203.128.0/19	100.42.48.0/23	103.150.187.121
100.0.114.17	103.109.101.105	101.248.0.0/15	100.42.50.0/23	103.27.202.127
100.0.127.86	103.163.218.11	102.196.96.0/19	100.42.52.0/22	103.41.177.43
100.0.13.109	103.196.37.111	102.211.224.0/19	100.42.56.0/23	104.131.1.24
100.0.166.88	103.236.201.88	102.212.224.0/19	100.42.58.0/23	104.207.130.126
100.0.183.143	103.251.167.10	103.100.168.0/22	100.42.60.0/22	104.232.35.45
100.0.193.122	103.251.167.20	103.13.140.0/22	101.0.232.0/21	107.189.13.164
100.0.221.72	103.251.167.21	103.134.226.0/23	101.0.232.0/22	107.189.2.146
100.0.241.26	103.253.24.18	103.14.208.0/22	101.0.240.0/20	108.53.208.157

Enrichment Procedure

Following the "Enrichment through GeoIP" process (Section 2.3.2) and utilizing "file 3", seventy-four distinct "Malicious Threat Intelligence Feeds" were utilized to process the "file 3" log file, creating a new file, "file 4". This process added more contextual information by merging the initial log file with threat intelligence data extracted from various sources identified in Section 2.3.3 (Definition and Explanation). Each IP was meticulously cross verified against each threat intelligence database. If an IP match was identified, a value of 1.0 was assigned to the respective field. Otherwise, it was marked with 0.0. The result was a comprehensive log file comprising the source IP and its presence or absence in any threat databases. This enhanced log file enabled swift and efficient cross-referencing and identifying potential threats.

The new log file ("file 4") consisted of a header followed by rows of data. The header incorporated the names of the fields, indicating the data source for each IP address. As an example, some of these headers are interpreted as follows:

- Source-IP: Denotes the source IP address, monitored for potential malicious activities.
- actor_indicators_apt28-ipv4-only.csv, actor_indicators_apt34-ipv4-only.csv: These headers indicate the source IP was linked with activities from specific APT groups, each represented by a unique number. These APT groups are recognized as cybercriminal or cyber-espionage entities.
- malicious-subnet-uceprotect-dnsbl-2.txt, malicious-subnet-misp-bro.txt, malicious-subnet-firehol-webserver.txt: These headers denote various threat intelligence databases, providing threat data at the subnet level.

- malicious-ip-abuse-urlhaus-recent.txt, malicious-ip-blocklist-bruteforcelogin.txt, malicious-ip-uceprotect-dnsbl-1.txt: These headers represent various threat intelligence sources or databases as well. However, they cater to specific IP-level threat data.

Each data row of the new log "file 4" incorporated a source IP address, followed by a series of binary indicators (0 or 1) for each threat intelligence source. A "0" indicated that the source IP address was not listed in the particular source, whereas a "1" signified its listing, denoting an association with malicious activity according to that source.

Final Preprocessing

In the culmination of the preprocessing stages described in Section 2.3.2. (Enrichment through GeoIP) and the section Enrichment Procedure, a consolidated output was derived, "file 4". This final output fused the enriched data from the GeoIP process and the augmented context obtained through integrating various threat intelligence feeds.

The final file was characterized by a comprehensive layout, including the geographical context associated with the source IP, relevant data on potential threat actor groups, and evidence of potential threats drawn from an extensive range of databases. This information was presented in a structured format, enabling efficient cross-referencing and swift identification of potential threats.

The enhanced IP data now contained geographical location details, affiliations with known cybercriminal entities, and an association, if any, with malicious activity as per various threat intelligence sources. Additionally, each source IP was meticulously checked and annotated according to its presence or absence in the threat databases, represented by binary indicators.

This final output, a result of rigorous preprocessing, served as the principal file for the study presented in this article. It represents a robust compilation of the enriched and transformed data, providing a foundation for meaningful analysis and insightful conclusions.

2.4. Validation

A range of validation techniques was employed in this study to ensure the accuracy and dependability of its findings. Post-exploratory data analysis, clustering algorithms, and anomaly detection methods were leveraged to identify the data's unique clusters or abnormal patterns. These techniques aid in discovering potential attack campaigns, recurring tactics, or behavioral patterns associated with cybercriminal activities [14,34].

A temporal analysis was undertaken, probing for variations in attack patterns, cycles in attack frequencies, and correlations with external incidents or threat reports. The derived outcomes were then cross-referenced with established sources of threat intelligence [8,14] to authenticate similarities or matches with available IOCs or known attack campaigns [13,23].

The validation of the data analysis results was also accomplished through comparison with extant literature. The trends, patterns, and transformations observed in this study were compared with findings from previous research [14,15,19]. This validation method offers an external reference point, allowing for comparing the study's findings with the prevailing understanding in the cybersecurity domain.

Ultimately, the findings were visualized and interpreted against the backdrop of the larger cybersecurity environment, providing a holistic understanding of the implications for threat mitigation, defensive strategies, or the creation of proactive measures to counter evolving cyberthreats [19,28].

By adhering to stringent validation protocols, the study aspires to yield reliable and valid findings that can enrich the ongoing cybersecurity discourse and assist in forming more robust and effective cybersecurity strategies [6,11,33].

3. Results

3.1. Introduction to Results

In cybersecurity, predicting and preventing cyberattacks rely heavily on a thorough understanding of adversarial Tactics, Techniques, and Procedures (TTPs). Cybersecurity emerges as a critical concern as we delve deeper into an increasingly interconnected digital world. The expanding threat landscape necessitates a deep understanding of attack patterns to formulate effective defense mechanisms. Here, honeypot logs serve as an invaluable asset. The study employs a honeypot log with a timespan of six years, covering daily cyberattack counts from October 2016 to September 2022.

This section presents the results from the systematic methodology applied to this dataset. The study's methodology involved the following steps: data preprocessing, exploratory data analysis, anomaly detection, temporal analysis, cross-referencing with established threat intelligence using OSINT, and visualization and reporting. Additionally, by applying advanced ML techniques and data analysis, strategies led to extracting many insights from the extensive dataset [8,14,15,18,25,30,32,35].

The stages resulted in significant findings, contributing to a comprehensive understanding of the patterns and strategies associated with cyberattacks. The results support H1 and shed light on the critical cyberattack trends and patterns (RQ1), indicating increased sophistication and target specificity over time. The findings highlight recurring techniques and bring to the fore emerging trends and tactics employed by cybercriminals [14,36,37].

The validation process substantiated the reliability and robustness of the findings, thereby bolstering the accuracy and credibility of the insights [3,14,19]. Furthermore, by cross-referencing the results with existing threat intelligence, the study was able to assess the evolution and relevance of the recorded cyberattacks [8,15,19].

Disseminating these results aims to contribute to the cybersecurity community significantly. Through these evidence-based insights, the study strives to guide the development of more effective defense strategies and proactive measures to counter advanced and targeted cyberthreats. The results contribute to the broader cybersecurity dialogue, underscoring empirical data's vital role in enhancing cyberdefense capabilities [1,19].

In conclusion, the results offer a comprehensive overview of the patterns, trends, and techniques prevalent in cybercriminal activities over six years. The results include focusing on the evolution in sophistication and target specificity, thus supporting Hypothesis (H1) and offering detailed answers to RQ1.

3.2. Data Collection and Preprocessing Results

The analysis was conducted on an extensive honeypot log file containing more than 100 million entries, recorded over six years from October 2016 to September 2022. This large volume of data provided a thorough view of cyberattacks, enabling the study to gain critical insights into these attacks' intensity, distribution, and patterns.

3.2.1. Summary and Descriptive Analysis

Throughout the six years, the average number of cyberattacks per day was 45,741. However, the variation around this mean was considerable, as shown by the standard deviation 58,788.5 (Table 2). This high standard deviation indicates a high level of dispersion in daily cyberattacks, with some days experiencing relatively fewer attacks and others having significantly more. Throughout the tracking period of 2191 days, 100,218,535 entries were documented in the honeypot system.

The highest number of cyberattacks recorded in a single day reached 888,203. This max value, coupled with a significant standard deviation, indicates the presence of days with extreme cyberattack counts, potentially corresponding to coordinated global cyberattacks or specific cyberevents. A day with zero attacks is rare, signifying the persistent nature of the threat landscape. Over the six years of log data, only 17 such instances were noted.

Table 2. Descriptive Statistics of Unique Daily Count.

Descriptive Statistics	
count	2191.000000
mean	45,741.001826
std	58,788.500082
min	0.000000
25%	16,037.000000
50%	28,447.000000
75%	58,430.500000
max	888,203.000000

The distribution of daily cyberattack counts shows significant skewness. The median number of attacks per day, 28,447, is substantially lower than the mean, suggesting a positively skewed distribution. That is, while most days experience a relatively moderate number of attacks, there are days with exceptionally high counts that push the average up.

Quartile ranges provide further insight into the distribution. The lower 25% of the data (the first quartile, Q1) shows that the number of attacks on a quarter of the days was 16,037 or fewer. The upper 25% of the data (the third quartile, Q3) indicates that on 25% of the days, there were 58,430.5 attacks or more. The interquartile range (Q3–Q1) stands at 42,393.5, showing a considerable spread in the middle 50% of the data.

3.2.2. Temporal Analysis

Temporal analysis of the cyberthreat landscape from October 2016 to September 2022 revealed that the volume of attacks could have been evenly distributed over the studied period. Figure 1 displays the occurrences by month. Significant peaks of activity were observed in July 2017 and October 2019. These periods signify instances of heightened threat activity. An upward trend in attack volumes was detected starting from the latter part of 2019. By 2021, it was commonplace for monthly attacks to surpass the 2 million mark, with February and August registering extraordinarily high attack counts of 3,252,302 and 3,491,482, respectively. Although there was a minor reduction towards the year's conclusion, the volume of attacks experienced a resurgence in 2022, with May and June recording over 3 million attacks each.

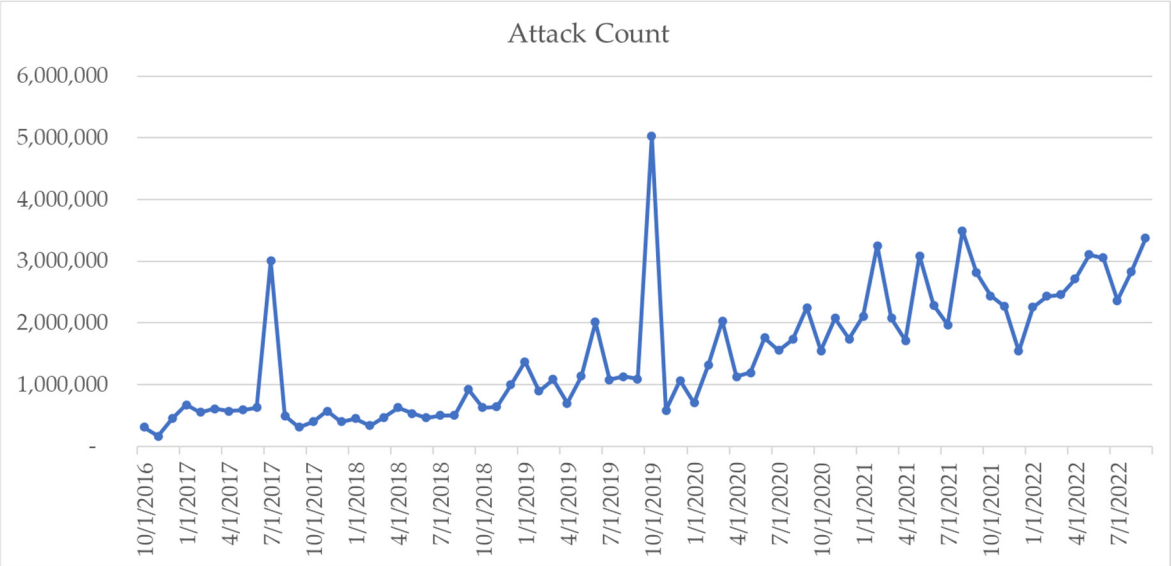


Figure 1. Count of Malicious Activity.

3.2.3. Correlation Analysis of Malicious Threat Intelligence Feed

The relationship between an array of cybersecurity-associated parameters was scrutinized via a correlation analysis, using Pearson correlation coefficients to characterize the magnitude and directionality of these associations. This study involved 2211 correlation calculations, comparing known threat intelligence sources (discussed in Section 2.3.3) to identify potential parallels or congruencies with established IOCs or recognized attack campaigns [8,15,23].

The analysis showed that 56 correlations were statistically meaningful, while 2155 were found not to hold any significant association. The spectrum of correlation values spanned from approximately 0.31 to a perfect 1. A pair of threat feeds, specifically "subnet-misp-bro" and "subnet-misp-ip-dst" (Table 3), revealed a perfect positive correlation of 1, signifying either identical or impeccably mirrored datasets. In addition, several pairs of files displayed high degrees of correlation, with coefficients of 0.987332677, 0.845672856, 0.822205602, 0.784329101, and 0.783025671. The corresponding *p*-values for all the meaningful pairs were 0, thus signifying the high statistical significance of these observed correlations.

Table 3. Correlation Calculations of Malicious Threat Intelligence Feed.

Malicious Threat Intelligence Feed—1	Malicious Threat Intelligence Feed—2	Correlation	*p*-Value
subnet-misp-bro	subnet-misp-ip-dst	1	0
ip-uceprotect-dnsbl-3	ip-uceprotect-dnsbl-2	1	0
ip-firehol-anonymous	ip-firehol-proxies	0.987332677	0
ip-dan-torlist-exit-ip	ip-dan-torlist	0.933264363	0
ip-blocklist-SSH	ip-blocklist	0.880540508	0
subnet-spamhaus-drop	subnet-snort-pulled-pork	0.845672856	0
subnet-snort-pulled-pork	subnet-firehol-spamhaus_drop	0.845672856	0
ip-firehol-webclient	ip-firehol-webserver	0.822205602	0
ip-misp-bro-ipv4	ip-misp-ip-dst-ipv4	0.784329101	0
subnet-firehol-anonymous	subnet-firehol-proxies	0.783025671	0

3.2.4. Geographic Analysis

The geographical analysis indicated a global distribution of cyberattacks, originating from six continents and 188 countries. North America recorded the highest number of entries, followed closely by Europe and Asia. Examining specific countries, the United States emerged as the primary source, contributing approximately 54.5% of the total entries. Russia, China, and The Netherlands were the most significant contributors, accounting for 17.4%, 9.4%, and 4.8% of the entries, respectively (Figure 2). Despite the broad distribution, none of the other countries exceeded 4% of the total entries individually (Figure 3). A concentrated activity was noted within the top 10 countries, contributing to 98.7% of the total entries. Meanwhile, a continental analysis revealed that North America, Europe, and Asia contributed 45.017%, 35.370%, and 22.180% of entries, respectively, while South America, Africa, Oceania, and Antarctica collectively accounted for less than 3%.

3.2.5. Threat Intelligence Analysis

Threat analysis of a unique source IP address dataset comprising 1,316,585 malicious entries revealed an attractive distribution. Utilizing the "Malicious Threat Intelligence Feeds" or "threat intelligence repositories", as discussed in Section 2.3.3, source IP addresses were categorized into zero (0) or non-zero (1). Whereas "0" represents a source IP address not listed in the threat intelligence repositories, and "1" indicates the source IP address was identified within the repositories. The total number of repositories consisted of seventy-four different and separate threat intelligence feeds. Sixty-five repositories were found to have positive matching source IP addresses from the dataset ranging from 1 match in one repository to as many as 581,115 matches in another.

Within the dataset of 1,316,585 unique source IP addresses, 699,543 (53.133%) unique source IP addresses aligned with the threat intelligence repositories, thus receiving a non-

zero count or a "1". In contrast, 46.867% or 617,042 entries received a zero count or a "0". Despite the lack of identification in the threat intelligence data, these zero-count entries represent significant threats due to behavior consistent with recognized and identified malicious conduct within the dataset. The analysis classified the entries under various categories, signifying distinct types of malicious activity or sources; examples of the number of matching repositories with the total number of matching source IP addresses for the respective repository can be found in Table 4. This study focuses on the source IP addresses receiving a non-zero count or a "1" as the result of a positive match within the threat intelligence repositories. Future research could focus on the non-positive matches.

Table 4. Example of Threat Intelligence Repositories Matches.

Malicious Threat Intelligence Feed	Matches	Total Source IP
subnet-uceprotect-dnsbl-3	581,115	44.138%
subnet-uceprotect-dnsbl-2	300,891	22.854%
subnet-firehol-webserver	91,983	6.986%
ip-misp-ip-dst-ipv4	41,701	3.167%
ip-misp-bro-ipv4	28,080	2.133%

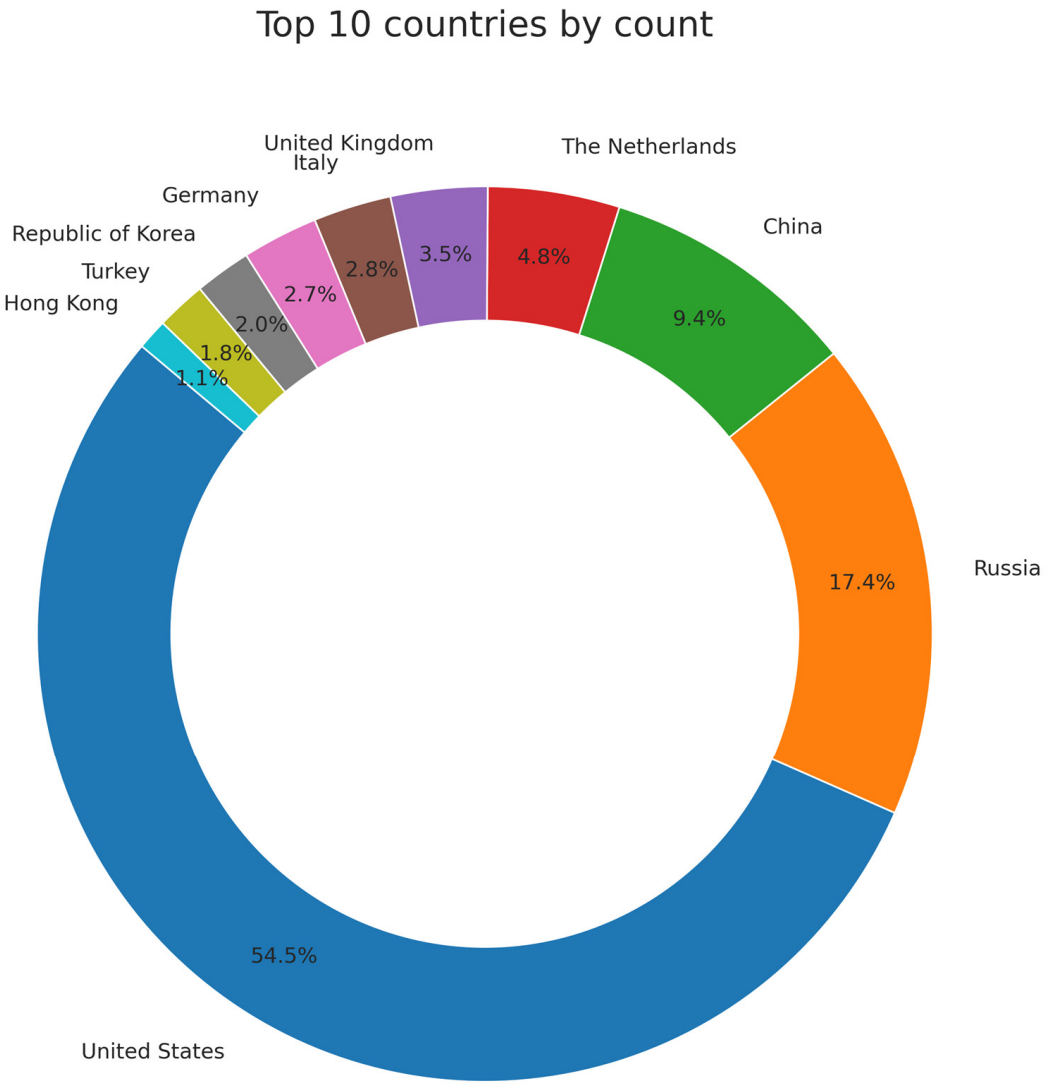


Figure 2. Top 10 Source Country.

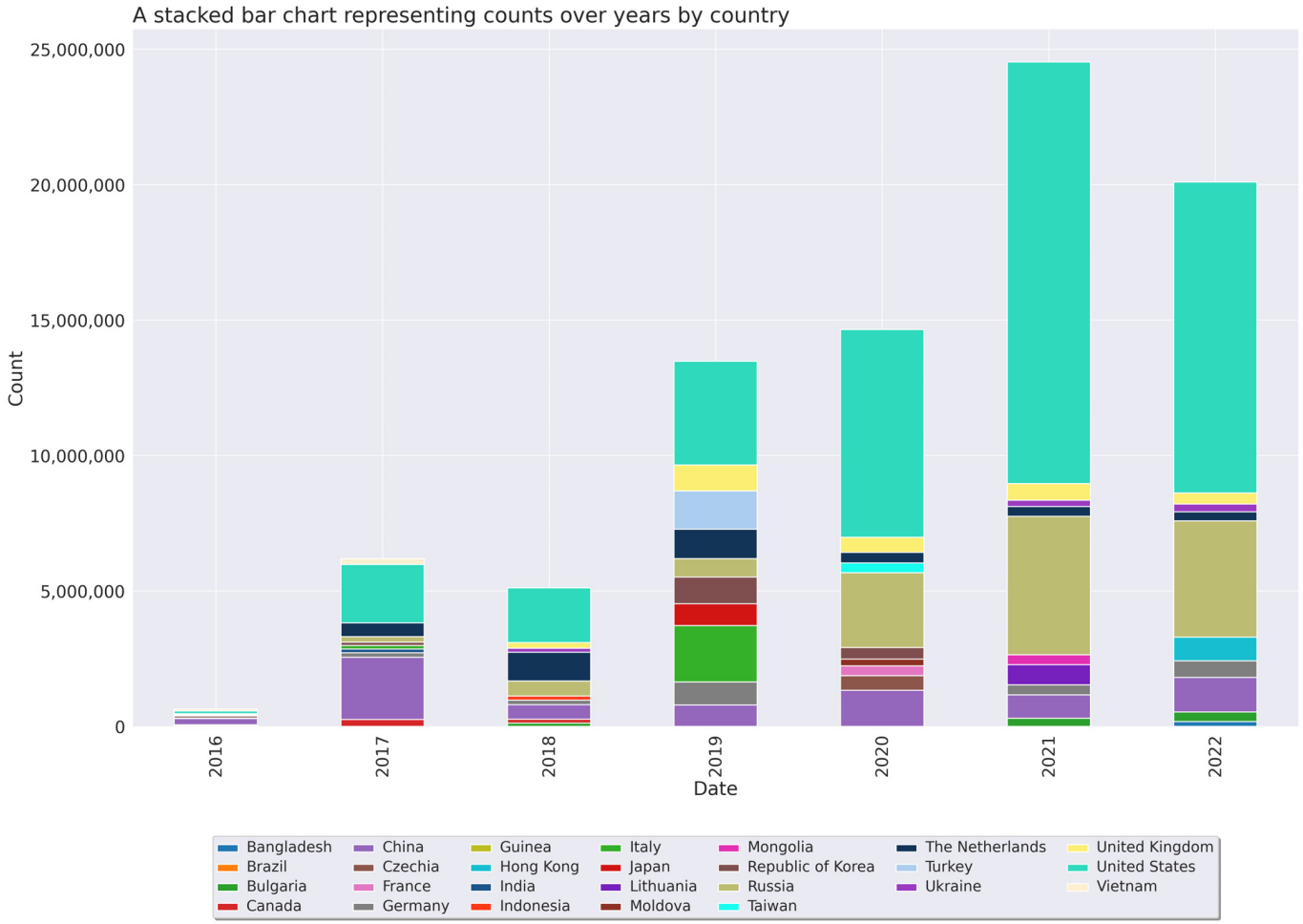


Figure 3. Top 10 Source Countries by Fiscal Year.

3.2.6. Source IP Address Analysis

The study analyzed the source IP addresses as a primary variable against threat intelligence data repositories. A total of 100,218,535 source IP address entries were collected over 2191 days. The log data encompassed entries from 1,316,585 unique source IP addresses, with the IP address 23.139.224.114 associated with the highest number of entries at 2,217,585 (Table 5). Further, the analysis revealed that the entries related to the top 20 unique source IP addresses amounted to 10,835,108, constituting approximately 10.81%. As previously discussed in Section 3.2.4, the highest number of entries geographically originated from North America, closely followed by Europe and Asia, with the United States, Russia, and China identified as the primary contributing nations.

Table 5. Overall Top 10 Source IP Addresses.

Source IP	Count	Percentage
23.139.224.114	2,217,585	2.215%
162.142.125.128	1,045,622	1.044%
100.27.42.150	758,851	0.758%
100.27.42.187	754,386	0.754%
100.27.42.157	693,224	0.693%
64.227.110.98	687,625	0.688%
92.63.197.18	677,060	0.677%
143.110.156.7	580,346	0.580%
161.35.232.85	569,259	0.569%
93.115.29.34	531,990	0.532%

3.2.7. Destination Ports Analysis

The log file indicated a total of 65,535 unique destination IP ports targeted. The three most targeted ports were 5900 (VNC server), 8 (ICMP), and 22 (SSH), which contributed 15.883%, 10.223%, and 4.459% of total entries, respectively. These ports are associated with remote control services and diagnostic tools, suggesting attackers tend to target remote access points and obtainting information from network diagnostic tools. The total entries accounted for by the top 20 ports were 66,741,317, constituting approximately 66.52%. A temporal analysis of network traffic from 2016 to 2022 indicated increased traffic over the years for specific ports, a decline for others, and significant traffic for certain ports in specific years only (Figure 4).

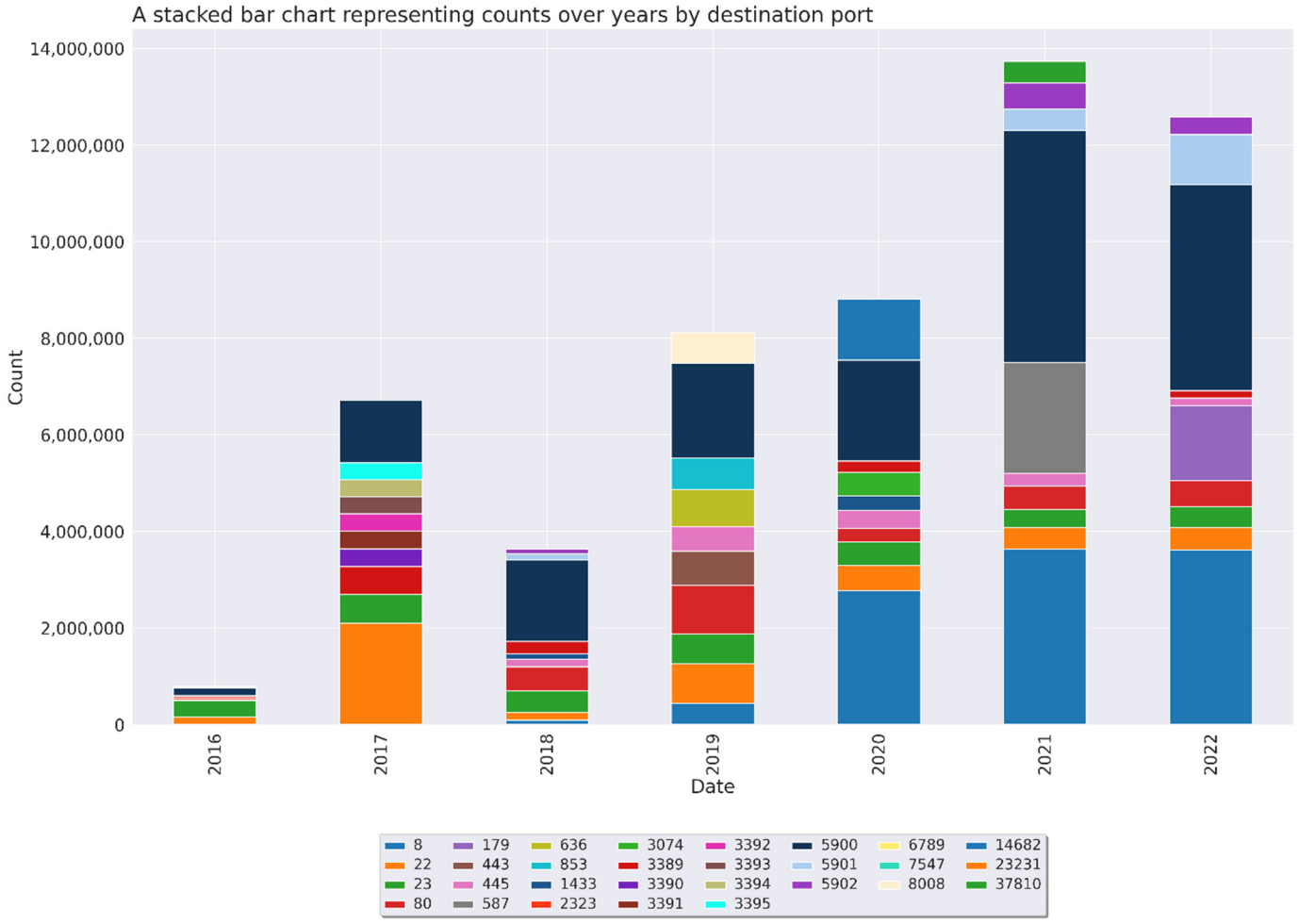


Figure 4. Top 10 Destination Ports by Fiscal Year.

3.2.8. Destination Services Analysis

The analysis of destination IP services disclosed the same results as the port analysis. Two hundred eighty-three unique services were targeted, with most entries tagged as "Unknown". As shown in Figure 5, the leading services were "Unknown", "VNC-Server", "ICMP–Echo-Request", "SSH", and "Telnet", contributing 54.0%, 16.8%, 10.9%, 4.8%, and 3.4% to the total entries, respectively.

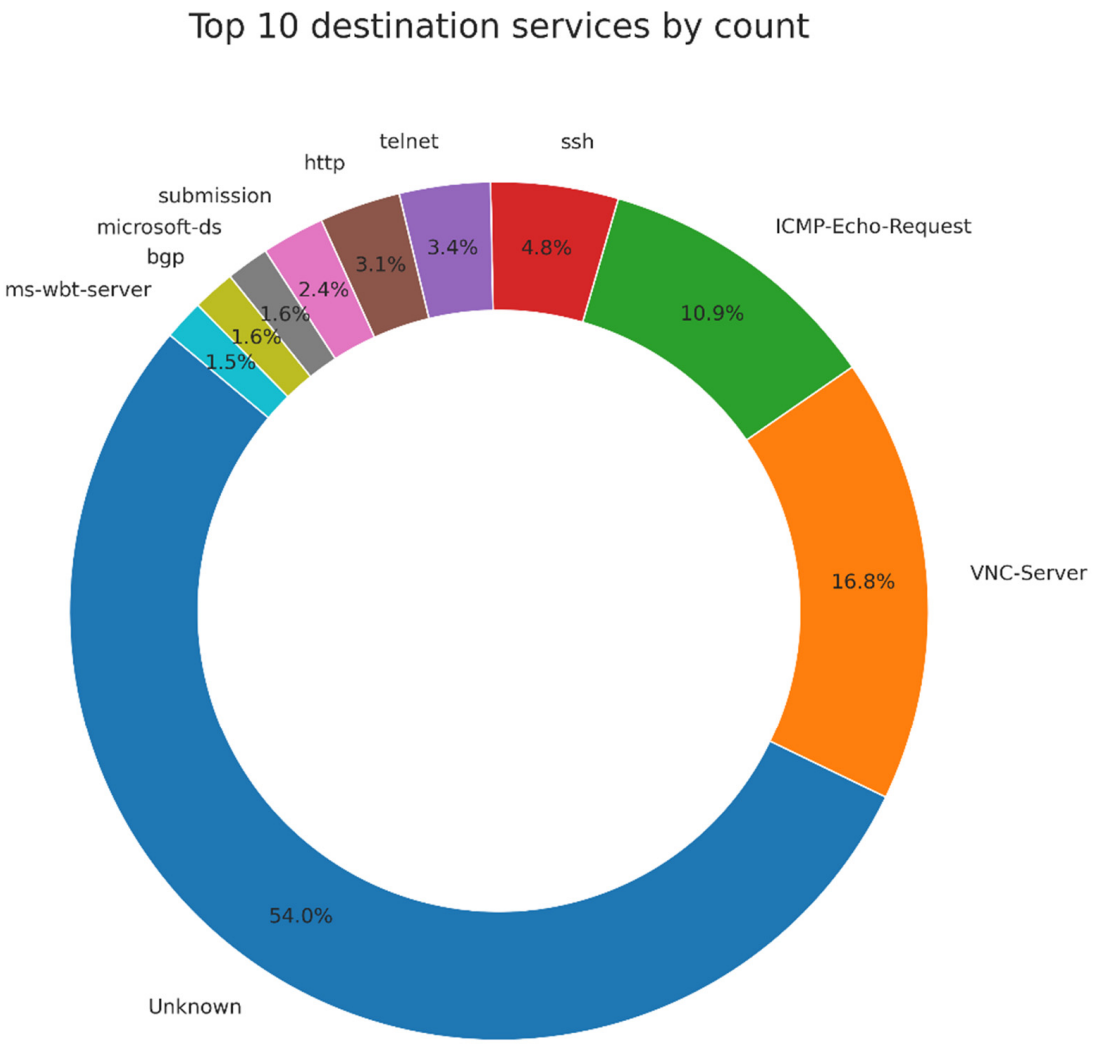


Figure 5. Pie—Top 10 Destination Services.

As shown in Figure 6, the results highlight a preference for remote access and network information amongst attackers and underscore the considerable proportion of attacks associated with the "Unknown" category. The "Unknown" service, contributing to nearly half of the total entries, presents a growing concern in cybersecurity. This category could comprise multiple services, including unconventional, newly devised, or obscure methods attackers use that are not easily classified or identifiable. The increased prevalence of "Unknown" signifies that attackers are innovating and employing methods that circumvent typical detection strategies. This increasingly opaque nature of attacks further complicates the task of cybersecurity, necessitating the development of more advanced and adaptive threat detection and prevention systems.

As shown in Table 6, a "Count Diff" column was added to each year for each destination service. This column indicates the annual change in traffic volume from year to year and provides evidence of the increasing complexity and sophistication of cyberattacks. A consistent yearly increase in the attack traffic was observed on the "VNC-Server" port, with the most substantial surge documented in 2021. On the other hand, services such as "SSH" and "Telnet" demonstrated more erratic patterns. The "ICMP–Echo-Request" service appeared in 2017 and has since been a consistent target, peaking in 2020. A new target, the "BGP" service, was observed in 2022.

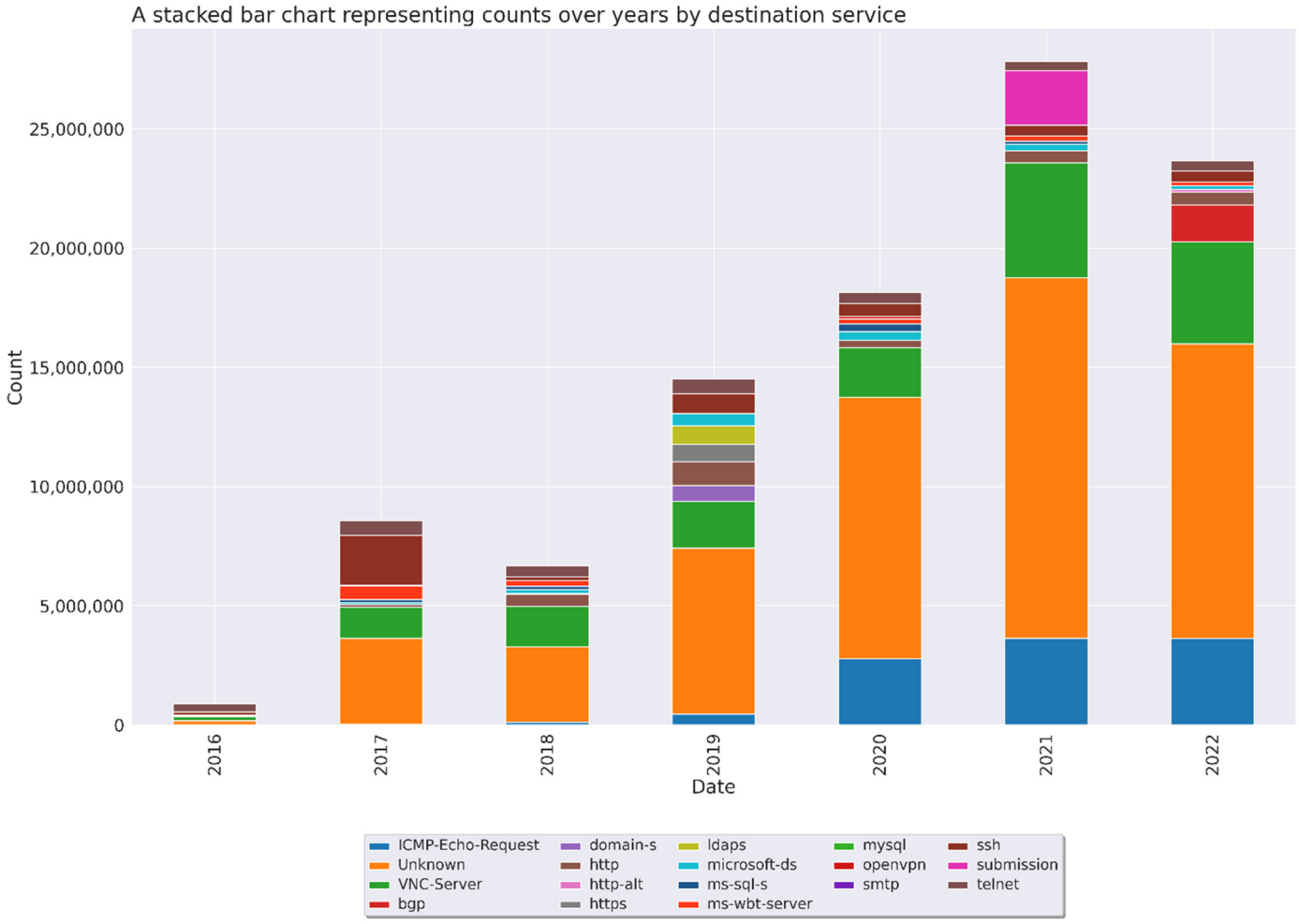


Figure 6. Bar—Top 10 Destination Services by Fiscal Year.

Table 6. Sample Data of the "Count Diff" Between Two Years.

Date	Destination Service	Count	Count Diff	Date	Destination Service	Count	Count Diff
2020	Unknown	10,967,734	3,992,397	2021	Unknown	15,111,843	4,144,109
2020	ICMP (8)	2,768,567	2,332,039	2021	VNC-Server	4,816,664	2,724,705
2020	VNC-Server	2,091,959	127,337	2021	ICMP (8)	3,633,149	864,582
2020	SSH	524,791	(299,721)	2021	submission	2,299,295	2,299,295
2020	Telnet	488,790	(135,510)	2021	http	504,602	214,473
2020	microsoft-ds	373,139	(136,318)	2021	SSH	441,571	(83,220)
2020	ms-sql-s	7593	185,504	2021	Telnet	382,065	(106,725)
2020	http	290,129	(189,936)	2021	microsoft-ds	283,560	(89,579)
2020	ms-wbt-server	230,567	(57,233)	2021	ms-wbt-server	221,742	(8825)
2020	openvpn	103,497	103,497	2021	ms-sql-s	126,171	(181,422)

3.2.9. Autonomous System Numbers and Names Analysis

A comprehensive analysis of Autonomous System (AS) Numbers and Names (ASNs) was performed on data from 2016 to 2022, identifying 21,110 unique AS. The top 20 organizations or ASNs made up approximately two-thirds of all entries, which indicates a high level of network activity originating from these particular source networks (Figure 7). The ASNs with the highest entries were 14,061 (DigitalOcean—United States), 14,618 (Amazon-AES—United States), and 16,509 (Amazon-02—Japan). Despite this, it is clarified that high entry numbers do not necessarily indicate the organizations' direct involvement in malicious activities but rather could reflect their large customer bases.

Among the source organizations identified in the study, 19,903 unique entities were found, including but not limited to DigitalOcean (United States), Amazon-AES (United States, Amazon-02 (Japan), Censys (United States), and Google Cloud (United States). A significant proportion of entries originated from networks based in China ("Chinanet" and "China Unicom 169 Backbone").

Temporal analysis revealed significant fluctuations in specific ASNs over time, including 4134, 14,061, 14,618, and 16,509. Further, a detailed cluster analysis showcased distinct clusters of ASNs, such as 134,176 and 208,091, which displayed significant growth in particular years.

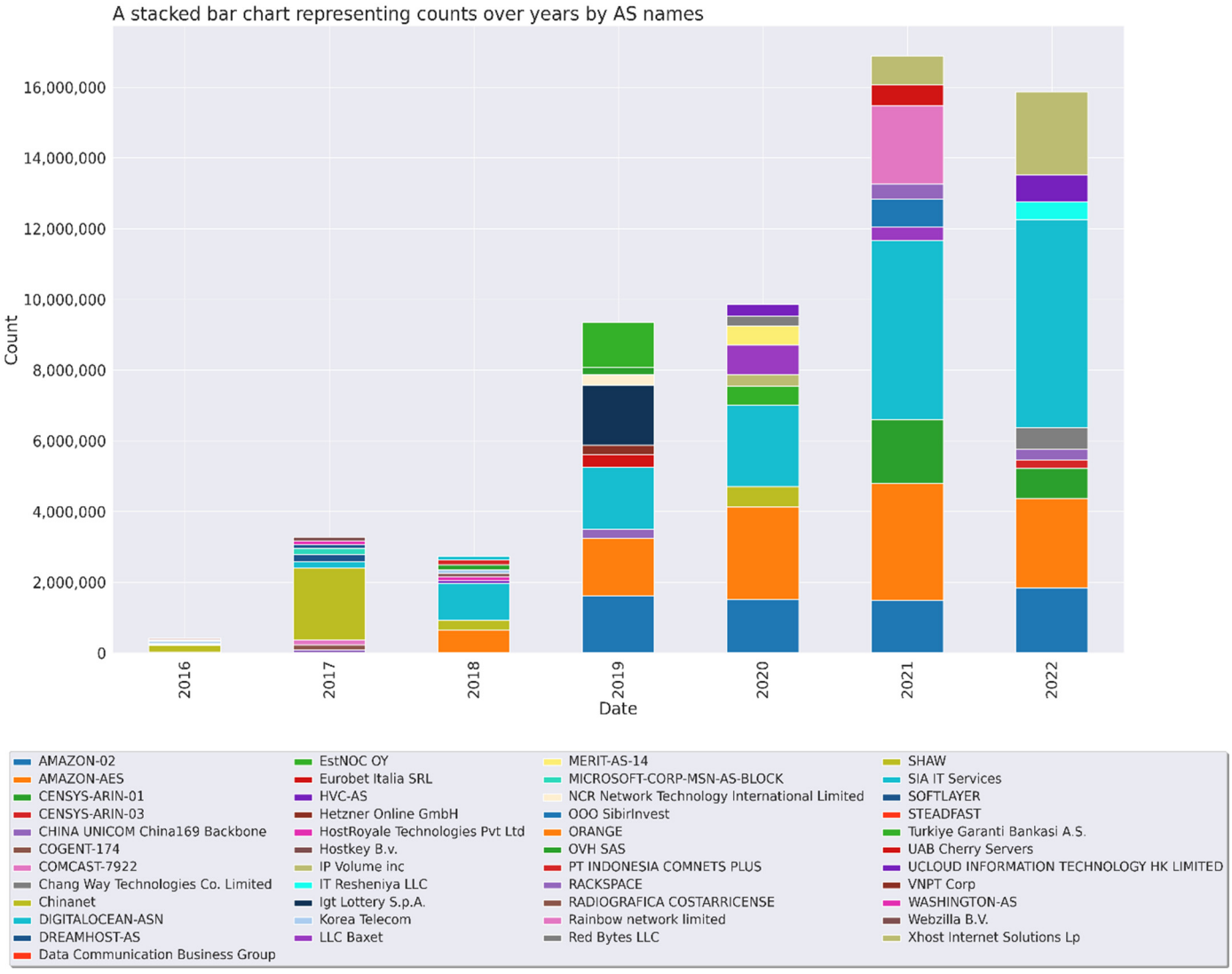


Figure 7. Top 10 ASNs by Fiscal Year.

3.2.10. Behavior Analysis

Behavior analysis was performed using an approach involving the calculation of a behavior score. The behavior score was defined based on the aggregated network traffic by source IP address, the respective AS number, and the organization to which the AS is registered. Table 7 showcases a subset of the dataset, demonstrating the relationship between source IP, country, ASN, and the corresponding behavior score. The behavior score was highest for "DigitalOcean" in the United States, "F3 Netze e.V". in Germany, and "CT-HangZhou-IDC" in China, indicating potentially high-risk anomalous behavior. Table 7 is sorted by behavior score and is an example of a partial list.

Table 7. Behavior Analysis (Partial Snapshot).

Source-IP	Source-Country	Source-AS-Org-Name	Behavior Score
185.220.103.9	United States	CALYX-AS	136
95.85.7.220	United States	DigitalOcean-ASN	136
60.191.87.89	China	CT-HangZhou-IDC	136
23.129.64.216	United States	EMERALD-ONION	120
171.25.193.80	Sweden	Foreningen for digitala fri-och rattigheter	120
199.249.230.87	United States	QUINTEX	120
92.255.85.9	Russia	Chang Way Technologies Co. Limited	120
83.229.82.236	The Netherlands	Kamatera Inc	120
66.102.248.138	United States	Chinanet	105
60.9.97.113	Mongolia	China Unicom China169 Backbone	105
89.190.159.189	South Africa	Alsycon B.V.	105

The distribution of behavior scores was further analyzed. Many entities exhibited a behavior score of 0, accounting for 617,042 instances, which matched the number of source IP addresses that received a "0" for not matching any of the threat intelligence repositories used in this study. However, behavior scores deviating from this baseline were scrutinized as potential anomalies. Notably, behavior scores of 1, 3, and 6 were registered for ASNs 368,742, 269,550, and 36,685 instances, respectively. As shown in Table 8, these observations suggest a gradient of potentially anomalous behavior with varying severity, as inferred from the behavior score.

Table 8. Behavior Analysis All Score Counts.

Behavior Score	Count
0	617,042
1	368,742
3	269,550
6	36,685
10	13,606
15	4606
21	2571
36	1280
28	1222
45	628
55	306
66	151
78	77
91	49
105	34
120	25
136	11

3.2.11. Clustering Analysis

This analysis conducted a comprehensive clustering procedure, applying a robust pipeline that integrated preprocessing and clustering stages. The pipeline was assembled employing Python's sklearn library, consisting of a column transformer (for preprocessing) and K-means clustering (for the clustering stage). The K-means algorithm was purposely initialized ten times to ensure the reliability of the resulting clusters.

The features selected for the clustering analysis encompassed "Date", "Count", and "AS-Number". After the pipeline and feature selection had been outlined, the pipeline was fitted and subsequently applied to transform the selected features of the data. This process yielded a scaled version of the data conducive to clustering analysis.

Upon transformation, the clustering procedure gave rise to three distinct clusters, each exhibiting varied distributions and frequencies of data points. Out of the total 21,740 data

points, the majority was absorbed by clusters 0 and 2, containing 10,786 and 10,929 data points, respectively, while cluster 1 was significantly smaller, with a mere 25 data points.

Further, the mean "Count" of data for each cluster was computed, unveiling considerable disparities among the clusters. The following was observed in the output data:

1. The average "Count" for clusters 0, 1, and 2 stood at 5217.48, 401,954.76, and 1410.81, respectively.
2. The averages suggested that cluster 1 comprised data points with a substantially larger "Count" relative to clusters 0 and 2.
3. Clusters 0 and 1 had no points in all the given months. All data points were assigned to cluster 2. These assignments were due to how the KMeans algorithm found the best fit for the data.
4. Few anomalies were detected, specifically in January and November 2017, and April 2018. These anomalies had a very high average count (55,489, 131,077, and 103,400, respectively), which could indicate high-volume attacks or significantly different malicious activity during these periods.
5. The average count of malicious activities for cluster 2 varied monthly, with the highest average in July 2017; this suggests that the volume and intensity of attacks can significantly vary over time.

Consequently, in the example shown in Table 9, the cluster labels resulting from the K-means algorithm were annexed to the original data frame, establishing a new "cluster" column. This addition supplied a precise label for each data point, indicating its belonging to a specific cluster (0, 1, or 2). This clustering analysis provided an informative overview of the data's inherent structure and paved the way for subsequent anomaly detection. Recall that the average "Count" for clusters 0, 1, and 2 stood at 5217.48, 401,954.76, and 1410.81, respectively.

Table 9. Example of Long-Term Trends with Clusters and Anomalies (21,740 Data Points).

Date	AS Number	Count	Cluster	Anomaly
29 September 2022	14,061	28,410	0	1
29 September 2022	398,108	20,011	0	1
29 September 2022	14,618	11,090	0	1
29 September 2022	212,512	8078	0	1
29 September 2022	398,324	3158	0	1
29 September 2022	50,867	2904	0	1
29 September 2022	57,523	1866	0	1
29 September 2022	22,612	1770	0	1
29 September 2022	55,960	1673	0	1
29 September 2022	270,764	1464	0	1
30 September 2022	14,061	20,767	0	1
30 September 2022	14,618	8110	0	1
30 September 2022	40,676	6998	0	1
30 September 2022	398,324	3406	0	1
30 September 2022	206,264	2893	0	1
30 September 2022	57,523	2307	0	1
30 September 2022	22,612	2021	0	1
30 September 2022	213,206	2013	0	1
30 September 2022	270,764	1580	0	1
30 September 2022	398,722	1490	0	1

The 0 in the anomaly count and the cluster count columns indicate that no anomalies or data points were assigned to that cluster in that particular month. Similarly, a value of 0 in the average count columns indicates that either no anomalies or data points were present in that cluster for that month; hence no average count could be computed.

The temporal distribution of "Count" within each cluster, gleaned from the summary analysis, displayed unique trends over time for each cluster, which will be further discussed in the next section. Initially, data points were predominantly in cluster 2; however, a progressive shift towards cluster 0 was observed over time. Data points in cluster 1 were found to be sporadic and carried higher "Count" values, leading to a greater mean for this cluster.

3.2.12. Anomaly Detection with Clustering

Post the clustering analysis (Section 3.2.11), the dataset was scrutinized for potential anomalies using the Isolation Forest algorithm. This algorithm was chosen for its efficacy in detecting outliers in high-dimensional datasets, with a contamination parameter of 0.01 set to indicate the expected proportion of outliers in the data. Upon fitting this algorithm to the scaled data, an "anomaly," as shown in the Table 9 column, was added to the original data frame to denote potential outliers.

The final output log file for the "anomaly detection with the clustering process" consisted of a header followed by rows of data. The header incorporated the names of the fields. These headers are interpreted as follows:

- Date: This column represents the month and year of the recorded data.
- cluster_0_count, cluster_1_count, cluster_2_count: These columns record the number of data points belonging to each cluster for each month. The code used K-Means clustering, which partitions data into K distinct, non-overlapping subsets (or clusters). In this case, K = 3 was used, so there are three clusters (0, 1, and 2).
- anomaly_count: This column shows the number of detected anomalies for each month. Anomalies are data points that are significantly different from the others. These could represent potential attacks or other abnormal behaviors.
- cluster_0_avg_count, cluster_1_avg_count, cluster_2_avg_count: These columns show the average count of malicious activity for each cluster in each month.
- anomaly_avg_count: This column shows the average count of malicious activity for the anomalies detected each month.

The analysis flagged 21,522 data points as non-anomalous and identified 218 as potential anomalies. The "Count" variable provided a distinguishing factor, as the mean "Count" for the anomalies was significantly higher than that for the non-anomalous data points. This comparison reaffirmed and validated the effectiveness of the applied anomaly detection method in identifying outliers based on the "Count" variable.

A discernible trend emerged when examining the temporal distribution of anomalies derived from the summary analysis. Anomalies were first observed in January 2017, initially appearing at a rate of one anomaly per month. However, the frequency of these anomalies increased over time, culminating in a peak of 17 anomalies in September 2021.

Figure 8 illustrates the effectiveness of the chosen clustering method in organizing the data and the anomaly detection method in flagging outliers. Both techniques provided valuable insights into the data's structure and behavior, with the Isolation Forest algorithm echoing the interest areas identified through clustering, thereby confirming the robustness of the chosen approach. These results are expected to contribute to further data analysis and enhance understanding in this field. Future work will delve deeper into these anomalies to unlock their full potential and implications. Recall that the analysis flagged 21,522 data points as non-anomalous and identified 218 as potential anomalies. The "Count" variable provided a distinguishing factor, as the mean "Count" for the anomalies was significantly higher than that for the non-anomalous data points.

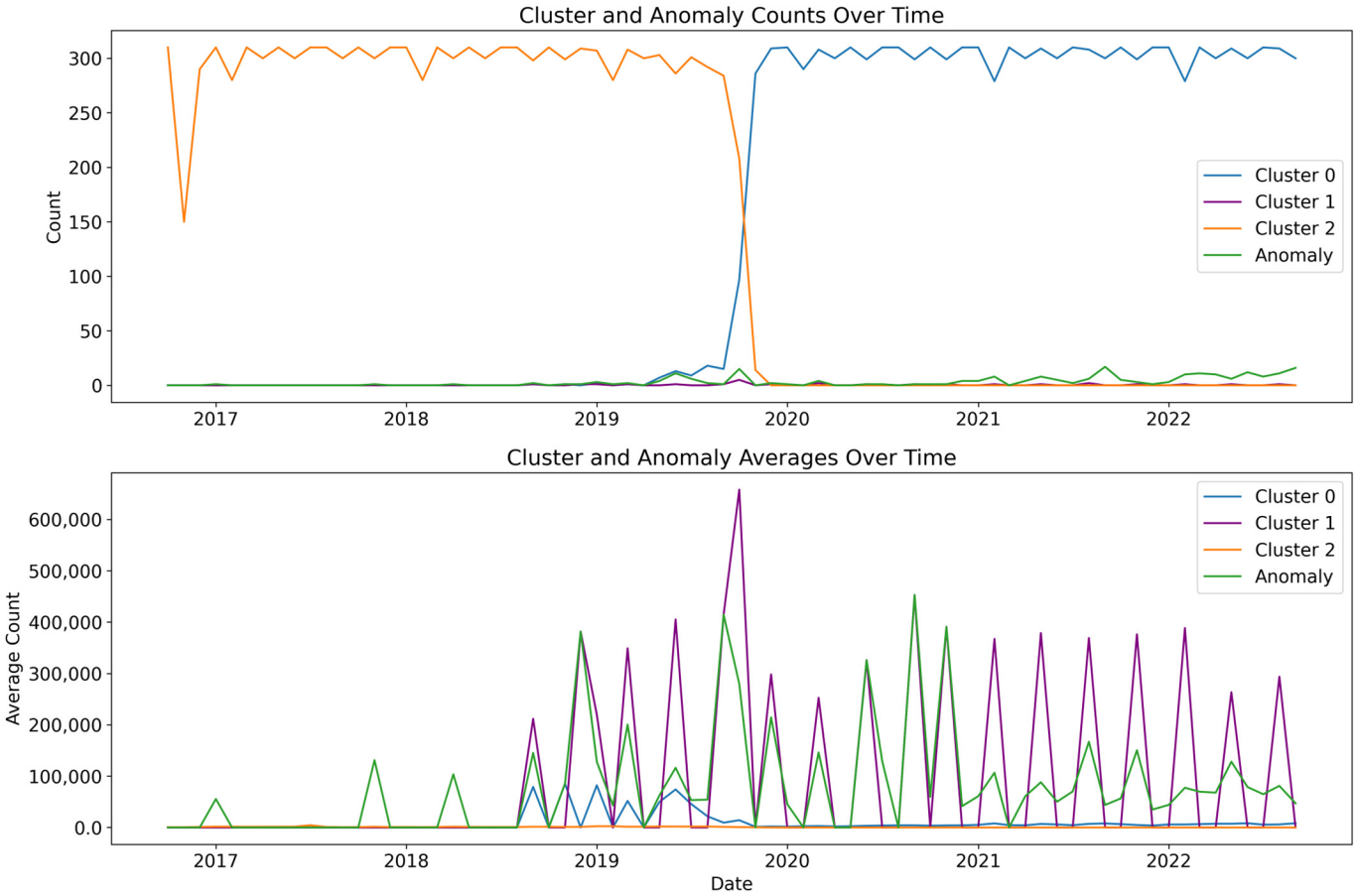


Figure 8. Cluster and Anomaly Counts and Averages.

3.3. Validation of Results

In the current study, validation of results was a pivotal aspect of the research methodology. It was instrumental in establishing the reliability and robustness of the cluster analysis and behavior scores deduced from the extensive network traffic data [38]. In this section, the process and outcomes of the validation stage are detailed, with a specific focus on the evolution of cyberattack tactics and techniques over time (RQ1) and the increasing sophistication and target specificity of these attacks Hypothesis (H1) [39].

The validation started with a detailed scrutiny of the results from the cluster analysis, taking inspiration from the methodology by Bagui et al. [38]. This research focused on various parameters, including source IP addresses, destination ports, and autonomous system numbers. These parameters were examined over different years, from 2016 to 2022, revealing clear trends and patterns that highlight an evolution in adversarial tactics and techniques, as suggested by RQ1. This observation aligns with the similar trends discussed by Bhardwaj et al., emphasizing the shift toward behavior-based models in threat detection [39].

The data were grouped into three distinct clusters, each representing a different category of network traffic behaviors. Significantly, these clusters demonstrated an increasing sophistication and target specificity of attacks, substantiating Hypothesis (H1). The evolution of these clusters over time provided further evidence of the changing nature of cyberthreats, resonating with Ntingi et al.'s findings that traditional threat detection methods might not be adequate in the face of exponentially advancing technology [40].

A vital component of the validation was evaluating the anomaly detection approach. This approach used a behavior score that ranged from 0 to 136. Instances with higher scores were flagged as potential anomalies, indicating potentially malicious network activities.

Importantly, these scores showed an increasing trend over the years, hinting at the rising sophistication and specificity of cyberattacks, again providing evidence for Hypothesis (H1). This anomaly detection approach was partially inspired by the behavior-based structured threat-hunting framework presented by Bhardwaj et al. [39].

Regarding geographical and autonomous system analysis, certain countries and autonomous systems consistently recorded high behavior scores. For instance, network traffic from Germany and the United States, linked with autonomous systems such as "DigitalOcean-ASN", "F3 Netze e.V.", and "Zwiebelfreunde e.V." scored highly on the behavior scale, indicating a shift towards targeted and sophisticated attacks. The analysis echoes the work of Davanian, who also found specific geographical regions and autonomous systems to be consistent sources of potentially malicious activities [41].

To summarize, the validation process confirmed the reliability of the cluster analysis and behavior-scoring approach and provided supporting evidence for the study's research question and hypothesis. The analysis and scoring methodologies effectively identified increasingly sophisticated and targeted anomalies in network traffic, reinforcing the findings from Bagui et al.'s comprehensive network traffic dataset based on the MITRE ATT&CK framework [38].

However, it is essential to note that the behavior score does not provide specific details about the nature of the detected anomaly, necessitating further investigation. The absence of these details leaves room for future research, particularly identifying the nature and type of evolving cyberattacks. The validation findings reinforce the claim that data analysis methods like clustering and behavior scoring can enhance anomaly detection in network traffic, bolstering network security [40].

3.4. Summary of Results

The present research systematically examined extensive network traffic data spanning six years from 2016 to 2022. The data included multiple parameters such as source IP addresses, destination ports, and autonomous system numbers. The results offered multifaceted insights into network traffic behaviors, anomalies, and their sources, reflecting an increasing trend of sophistication and specificity, supporting Hypothesis (H1).

A cluster analysis of the data yielded three distinct clusters, each representing different network behaviors, thereby underscoring the diversity and complexity inherent in network traffic patterns [40]. Interestingly, an observed progression towards sophisticated and targeted anomalies was seen in these clusters, lending support to Hypothesis (H1) [41]. These anomalies were subsequently classified based on these clusters, enabling a comprehensive and structured approach to anomaly detection [42].

Time-series data analysis revealed temporal patterns in attack counts, pinpointing periods of notable anomaly. For instance, significant spikes in attack counts were identified in July 2017, December 2018, and October 2019, suggesting an evolution in adversarial tactics and techniques. These patterns not only reinforced Hypothesis (H1) but also guided proactive network security measures.

A behavior-scoring methodology provided a quantitative metric to identify potential anomalies. Each instance of network traffic has attributed a score from 0 to 136, with higher scores suggesting potential anomalies. Validation of these scores revealed their effectiveness as a reliable indicator of abnormal behavior, thereby substantiating the rising trend in scores over the years and the corresponding increase in attack sophistication [43].

Geographical and autonomous system analysis illuminated the sources of anomalies. Specifically, network traffic originating from specific countries (e.g., Germany and the United States) and associated with autonomous systems like "DigitalOcean-ASN", "F3 Netze e.V.", and "Zwiebelfreunde e.V." consistently showed higher behavior scores [38,43]. These elevated scores hint at a more significant potential for anomalies and, hence, at the growing sophistication and specificity of the originating attacks, supporting H1 [40,43].

The study's findings were diverse and enlightening, underscoring the value of applying data analysis methodologies to network traffic data [39,43]. The employed techniques—

clustering, time-series analysis, and behavior scoring—proved effective for anomaly detection in network traffic, a critical aspect of enhancing network security [38,41]. However, it was noted that additional research is needed to understand the exact nature of the detected anomalies, a facet not fully captured by the behavior score [43]. The ongoing evolution and increasing cyberattack sophistication validated the study's findings and emphasized the need for ongoing research and proactive defense strategies [39,40].

4. Discussion Section

4.1. Introduction

The outcomes of this analysis offer significant insights into the evolving global cyberthreat landscape, underscoring RQ1 concerning notable trends and patterns in cyberattacks during the period under study. The observed trends confirm Hypothesis (H1), revealing the growing complexity and target-specific nature of cyberattacks over time [39].

However, numerous entries from specific countries or IPs may not inherently indicate malicious intent. Still, such data furnish crucial knowledge about areas of high cyberactivity. This information can guide the development of geographically adapted cybersecurity strategies, addressing the escalating sophistication and target specificity of cyberattacks, as highlighted in Hypothesis (H1). A pertinent observation by Ntingi et al. (2020) is worth noting here, asserting the need for a more proactive approach, such as Cyberthreat Hunting, due to technological advances and global interconnectivity [40].

It must be noted, however, that the locations of cybercriminals often remain concealed, potentially leading to geographic data inaccuracies. The geographic data might not accurately reflect the cyberattackers' authentic origin [38].

The patterns and evolving strategies identified in the analysis are consistent with the aspects mentioned in RQ1 and Hypothesis (H1). The analysis further unveils the perpetrators' increasingly intricate tactics to target various services, including those used for remote access and network diagnostics. These strategic advances align with the trends identified in the research question and hypothesis, providing concrete evidence of the continually evolving cyberthreat landscape [43].

Acknowledging that these emerging trends and patterns can inform cybersecurity strategies, policymaking, and resource distribution is essential. The findings indicate the necessity of developing more complex and forward-thinking measures to counter the increasing sophistication and target specificity of cyberattacks. The findings align with the observations made by Bhardwaj et al., advocating for a shift from traditional, signature-based reactive threat detection solutions to a proactive, behavior-based approach [39].

Furthermore, the work by Davanian provides valuable insights, particularly regarding intrusion detection techniques and IoT malware behavior [41]. These insights could serve as valuable benchmarks for ongoing research in this area. His findings about the behavior of bots and Command and Control servers over time could provide an exciting approach to tracking the evolution of cyberattack tactics and techniques [41].

Using comprehensive network traffic datasets such as UWF-ZeekData22, as proposed by Bagui et al., could be beneficial for future similar research [38]. Such datasets can identify attack traffic and detect adversary behavior leading to an attack and create user profiles of groups intending to perform attacks. The dataset's public availability also allows it to be used as a benchmark for this and other research [38].

Lastly, the systematic mapping study of deep learning techniques by Torre et al. for detecting cybersecurity attacks provides a broader context. It could be a benchmark for this study as well [42]. Despite the focus of the current study on network traffic data and behavior scoring, it can be worthwhile to discuss alternative detection methodologies in future research [42].

4.2. Interpretation of Results

The interpretation of the acquired results from the conducted analysis provides an intriguing understanding of the complexity and diversity of network traffic behavior. This

understanding aligns with RQ1 by illustrating principal trends and patterns in cyberattacks throughout the examined period. The results support Hypothesis (H1), affirming the escalation in sophistication and targeted cyberattack approach.

Classifying network traffic data into distinct clusters presents variability in network behavior patterns. These patterns are fundamental to understanding when developing sturdy security measures to counter increasingly sophisticated threats [41]. Each cluster exemplifies unique network characteristics, necessitating specialized preventative and responsive measures to maintain network security amidst growing attack specificity effectively.

The time-series analysis of the data captures the temporal patterns in attack counts, identifying periods of unusual activity or anomalies. The unusual activity could be attributed to the increasing cyberattack sophistication, as suggested in H1 [43]. Notably, the periods with spiked attack counts, specifically in July 2017, December 2018, and October 2019, emphasize the necessity for a temporal approach to network security as the techniques evolve in alignment with RQ1.

The behavior score, ranging from 0 to 136, is a quantifiable measure of potential anomalies [39]. The behavior score is a tool to quantify the increasing sophistication and target specificity of cyberattacks. Validation of these scores has accentuated their effectiveness as reliable indicators of abnormal behavior.

Geographical and autonomous system data are critical to comprehend the sources of network anomalies [38]. The higher frequency of abnormalities sourced from the United States and Germany and specific autonomous systems, namely "DigitalOcean-ASN", "F3 Netze e.V.", and "Zwiebelfreunde e.V." suggests that these areas and systems need close monitoring. This suggestion is due to the evolving nature and increasing cyberattack specificity, as outlined in RQ1 and Hypothesis (H1).

Despite these enlightening discoveries on the nature and origins of network anomalies, it is significant to acknowledge that the behavior score indicates the likelihood of abnormalities but does not identify the exact type or severity of the anomaly. This revelation resonates with Hypothesis (H1)'s proposition of escalating sophistication, considering that emerging attacks may diverge from recognized patterns [39]. As such, further research should aim to augment the current methods with procedures to identify the exact nature and potential implications of detected anomalies, particularly as threats evolve to be more intricate and targeted [41].

In conclusion, interpreting these results underscores the multifaceted nature of network traffic and the necessity for a comprehensive approach to ensuring network security. As suggested by RQ1 and Hypothesis (H1), the dynamic nature of cyberthreats necessitates a multi-pronged approach to counter them. This approach integrates temporal, geographical, and autonomous system data along with a quantitative measure of behavior [40]. These elements should all be considered to effectively identify and address network anomalies in an ever-evolving threat landscape.

4.3. Data Collection and Preprocessing

The comprehensive and rigorous data collection and preprocessing procedures undertaken in this study significantly enhanced the reliability and validity of the findings. Before the analysis, the data underwent meticulous cleaning, normalization, and transformation processes to ensure consistency and validity. Feature extraction and data transformation techniques were applied to the dataset, playing a pivotal role in extracting relevant information and ensuring the overall quality of the study's results. The careful data collection and preprocessing procedures enhanced the preparation of the dataset for subsequent analysis, contributing to the robustness of the study's outcomes.

The findings from this dataset reveal a complex and multifaceted cyberthreat landscape. The striking disparities in the origin of attacks highlight the global nature of the cyberthreat, pointing toward the need for enhanced international cooperation and coordination in addressing cyberthreats. However, it is also important to note that these disparities may be influenced by various factors, including the digital infrastructure, poli-

cies, and practices in different regions, as well as the ability of attackers to disguise their actual location.

These insights underscore the need for continuous monitoring and analysis of cyberactivities and for developing effective and adaptive strategies to mitigate cyberthreats. This study demonstrates the advantage of such comprehensive data collection and preprocessing efforts in generating critical insights that can inform policy and practice in cybersecurity.

4.3.1. Descriptive Analysis

The observed daily frequency of approximately 45,741 entries and the peak of 888,203 attacks in a single day reveal the scale and intensity of cyberthreats. The sporadic non-attack days, such as 16 November 2016, suggest periods of relative calm or a shift in attack strategies. These patterns underscore the dynamic nature of the cyberthreat landscape, requiring constant vigilance and adaptive responses.

The analysis underscores the erratic and volatile nature of cyberattacks, with daily counts varying wildly over the six years. The high degree of variation and the skewed distribution highlights the challenge of predicting and preparing for cyberthreats. Days with no recorded attacks are rare (17 out of 2191 days), reinforcing the constant nature of the cyberthreat landscape.

The marked distribution disparity points towards the global nature of cyberthreats, highlighting the necessity for international cooperation to mitigate these threats effectively. However, it is essential to remember that these distribution disparities might only partially represent the actual origin of the attacks, as cybercriminals often obscure their actual locations.

The descriptive analysis of the honeypot log presents a quantitative understanding of the cyberthreat landscape. The observed distribution disparities, peak activity, and periods of calm comprehensively depict cyberactivities. This study lays the groundwork for further analysis and interpretation of cyberthreats, emphasizing the importance of data-driven strategies to strengthen cybersecurity. The findings underscore the dynamic and complex nature of the cyberthreat landscape, reiterating the need for robust and adaptive cybersecurity measures informed by meticulous data analysis.

4.3.2. Temporal Analysis

The temporal analysis yielded a critical understanding of the cyclical trends in cyberattacks. The marked peaks in July 2017 and October 2019, followed by an overall increase in attack volumes from late 2019 onwards, point to an evolving and escalating cyberthreat landscape. These patterns suggest that cyberthreats are becoming more sophisticated and targeted, aligning with the initial Hypothesis (H1) that cyberattacks show a marked increase in sophistication and target specificity over time.

However, it is essential to consider the possibility of attack automation and an overall increase in Internet activity contributing to these high volumes. The variations in attack volumes could also indicate changing attacker tactics, advancements in detection methods, or the influence of global events. Consequently, these temporal trends necessitate ongoing evaluation to adapt and update cybersecurity measures in response to the evolving threat landscape.

The findings emphasize the importance of continual monitoring, evolution, and adaptation of cybersecurity strategies to detect and mitigate threats effectively. The study substantiates the growing significance of data-driven approaches to understanding and addressing the complexities of cyberthreats in the evolving digital era.

4.3.3. Correlation Analysis

The moderate to high correlations observed between the source AS numbers, corporate names, and numerous other indicators of malicious Internet activity suggest potential associations within the parameters studied. Such meaningful relationships may assist in predicting and identifying malicious activity based on known patterns. However, it must

be emphasized that correlation does not imply causation, thereby necessitating further examination to ascertain causal relationships between these variables.

In interpreting these correlations, one could hypothesize that attackers may utilize specific AS numbers, as indicated by the high correlations. However, additional factors such as the nature of the organization and its Internet traffic, the network infrastructure, and other contextual factors could influence these correlations. Therefore, considering these variables in future investigations would be crucial to validate better and comprehend the observed correlations.

The study underscores the necessity for a cautious interpretation of these correlations and the importance of further research to establish causal links. These findings highlight the potential of data-driven, statistical approaches to augment understanding and predict cyberthreats, contributing to more efficient and proactive cybersecurity strategies.

4.3.4. Geographic Analysis

The geographic distribution of cyberattacks offers crucial insights into the patterns of malicious cyberactivity. The significant fraction of cyberactivities originating from the United States, Russia, and China could indicate several factors, including technological advancement, economic influence, and geopolitical relevance. However, it is worth considering that cybercriminals frequently mask their precise location, which could skew the geographic data. Furthermore, the high concentration of cyberactivity within the top 20 countries might reflect their technological infrastructure and international standing. Such insights could be instrumental in shaping geographically precise cybersecurity policies and strategies. However, future studies should address the potential discrepancies resulting from attackers' masking of specific locations.

These findings emphasize the global nature of cyberthreats and highlight the importance of international cooperation and strategy development in cybersecurity. However, it is crucial to note the potential for location obfuscation by attackers, indicating the need for additional corroborative strategies to trace the origins of cyberthreats accurately. These threats' complex and international nature necessitate a multifaceted and global response.

4.3.5. Threat Analysis

The threat analysis presented in the study underscores the complexity and diversity of the cyberthreat landscape. A substantial number of unidentified threats (zero-count entries) emphasize the continual evolution of cyberthreats and the limitations of current threat intelligence repositories in capturing the complete range of malicious activity. The prominence of specific categories in the non-zero count entries signifies the prevalence of malicious activities or sources, providing valuable insights for devising targeted defense strategies. However, it is also critical to note the importance of minor categories, which, although constituting a smaller portion of the dataset, may represent emerging or less common threat vectors that warrant further exploration.

The significant number of unidentifiable threats reiterates the need to continuously enhance threat intelligence repositories and adopt adaptive, multifaceted cyberdefense strategies. The study's findings highlight the importance of ongoing research to understand the rapidly changing nature of cyberthreats and develop effective strategies to counter them.

4.3.6. Source IP Address Analysis

The study highlights the importance of scrutinizing the source IP address variable in understanding the origins and patterns of cyberattacks. The findings suggest concentrated sources of attacks from specific IP addresses and ASNs, pointing towards the potential utilization of botnets or centralized attack mechanisms. Notably, a significant percentage of entries were linked to the top 20 IP addresses, suggesting a concentrated nature of cyberthreats. The findings indicate a need for increased vigilance even in environments perceived to be trustworthy, particularly considering the predominant utilization of reputable cloud services as attack vectors. Understanding the dispersion and concentration of attacks

from individual source IPs informs the development of targeted defense mechanisms and fosters international collaboration to counter cybercrime effectively.

When cross-referenced with threat intelligence data repositories, the comprehensive analysis of source IP addresses revealed critical insights into the distribution of cyberthreats. The study reaffirms the necessity of an exhaustive analysis of source IP addresses to comprehend cyberattack patterns and develop effective threat detection and prevention strategies. By fostering international collaboration and sharing these insights, this approach contributes to the broader cybersecurity field's capacity to navigate the myriad cybersecurity challenges.

4.3.7. Destination Ports Analysis

The study's findings suggest an increasing sophistication and targeted approach to cyberattacks. The high prevalence of attacks on services like the VNC-Server (port 5900) that require more sophisticated attack vectors compared to standard ports such as HTTP (443) or SSH (22) reinforces this observation. The data points to a high concentration of attacks from specific IP addresses and ASNs, implying the potential use of botnets or centralized attack mechanisms. Using reputable cloud services to initiate attacks emphasizes the need for advanced security measures.

The "count_diff" data provides a dynamic perspective on the changes in network traffic over the years. Cyberattacks have become more targeted and sophisticated, with changing preferences for specific ports across different years. The fact that ports such as 5900 and 8 show a marked increase in traffic points to shifting attacker strategies. Conversely, a decrease in traffic for port 22 may suggest changes in the targeted systems' security measures or network configurations. Such insights could be leveraged in future cybersecurity studies and provide network administrators with crucial information to bolster network security measures. Therefore, the study offers a vital understanding of the cyberthreat landscape, underlining the need for continuous vigilance and adaptability in response to changing cyberthreats.

4.3.8. Destination Services Analysis

The study's results suggested an increased focus on less known or difficult-to-categorize services, indicative of a rise in the complexity and sophistication of cyberattacks. This finding aligns with the initial hypothesis. A consistent pattern of annual increases in attacks was noted for certain services such as "ICMP-Echo-Request", "Unknown", and "VNC-Server". In contrast, other services, such as "BGP" and "Domain-s", were only recorded in specific years.

The "Cluster" column, introduced through a KMeans clustering algorithm, provided additional depth to the analysis. It grouped destination services into clusters based on similarity, revealing distinct patterns for services like "Unknown", "VNC-Server", "ICMP-Echo-Request", and "SSH".

The analysis of "Destination Services" and the incorporation of the "count_diff" data and KMeans clustering painted a comprehensive picture of the evolving nature and complexity of cyberattacks. The analysis of destination IP services revealed a diverse range of targeted services. It demonstrated a marked increase in attacks on less known or harder-to-categorize services, indicative of increased complexity and sophistication of cyberattacks. These results are of immense value to network administrators and security professionals, providing vital insights for developing and reinforcing robust cybersecurity measures in response to the evolving threat landscape.

4.3.9. Autonomous System Numbers and Names Analysis

Despite the significant network activity linked to entities such as DigitalOcean, Amazon-AES, and Amazon-02, it is crucial to understand that these organizations' high entry numbers do not necessarily signify direct involvement in malicious activities. These

numbers reflect the large customer bases of these organizations, which could include users exploiting these services for nefarious activities.

Temporal trends demonstrate the ever-changing nature of the cyberthreat landscape. The fluctuations observed in specific ASNs over the years highlight the need for continuous monitoring and updating of cybersecurity measures to match the evolving nature of threats. Furthermore, the cluster analysis of ASNs offered more profound insights into the patterns of malicious network activity, indicating the changing landscape of cyberthreats.

The analysis of ASNs revealed distinct patterns of network activity linked to malicious intent, with significant variations across different ASNs and years. The findings emphasized the critical role of robust cybersecurity measures and continuous cyberthreat analysis in understanding and combating these evolving threats. By shedding light on the temporal behavior and clustering characteristics of ASNs, this analysis provides insights for future research in this area, thereby contributing to a broader understanding of cyberthreats and strengthening the defenses against them.

4.3.10. Behavior Analysis

As a metric, the behavior score demonstrated its potential in discerning anomalous from expected network behavior. This approach leverages the inherent structure of the Internet, employing AS numbers and organizations as critical factors in behavior analysis.

In the context of cyberthreat intelligence, these results highlight behavioral patterns' significant role in network traffic analysis. Countries like the United States and Germany exhibited higher behavior scores through their AS numbers and organizations, signaling potential security threats. Notably, these countries are significant Internet nodes, reinforcing the necessity of vigilant cybersecurity measures in these regions.

However, it is essential to consider that a higher behavior score may not directly correspond to malicious intent. Network traffic can exhibit strange behavior for several reasons, such as configuration changes, software updates, or non-standard user behavior. Therefore, these results should be interpreted with caution and need to be corroborated with additional data or context.

This study shed light on the potential of using behavior scores as an effective tool for anomaly detection in network traffic. The high behavior scores associated with specific AS numbers and organizations emphasize the need for rigorous and continuous monitoring of these entities. These findings and the distribution of behavior scores offer valuable insights for cybersecurity practitioners in their ongoing efforts to detect, mitigate and prevent cyberthreats.

While the study offers promising results, future work should focus on refining the behavior score by incorporating more diverse factors. Such enhancements will contribute to a decrease in false positives and improve the precision of the anomaly detection process. Additionally, further research is required to understand the reasons behind the elevated behavior scores observed for certain entities. Understanding these anomalies more deeply will facilitate the development of more effective threat intelligence strategies.

4.3.11. Clustering Analysis

In the presented analysis, a rigorous and comprehensive clustering approach was employed. This approach was underscored by a robust pipeline integrating preprocessing and clustering stages, providing a systematic and reproducible way to manage the complexity of the data. The pipeline was constructed using Python's sklearn library and composed of a column transformer and K-means clustering, with the K-means algorithm being purposefully initialized ten times to ensure the reliability and stability of the resulting clusters.

Key features, including "Date", "Count", and "AS-Number", were selected for the analysis. These features were processed by the pipeline, resulting in a scaled version of the data. The transformation of the data was not merely a technical procedure but rather a vital step that prepared the data for an effective clustering analysis.

The clustering procedure identified three distinct clusters with different characteristics and frequencies of data points. It was observed that clusters 0 and 2 contained the majority of the data points, while cluster 1 had significantly fewer data points. The disparities in the size of the clusters raised interesting questions about the underlying structure of the data and the significance of cluster 1, which may require further investigation.

Moreover, the computation of the mean "Count" for each cluster revealed considerable variations among them. This finding highlighted the heterogeneous nature of the data and provided insight into the potential significance of the different clusters.

The temporal distribution of "Count" within each cluster was also examined, revealing unique patterns and shifts over time. The shift from cluster 2 towards cluster 0 suggested dynamic changes in the data over time, which could be an interesting area for future study. The sporadic and high "Count" values in cluster 1 were intriguing and may suggest anomalies or rare events in the data.

The addition of cluster labels to the original data frame provided an insightful layer of information. Each data point was tagged with a precise label indicating its cluster affiliation, which could be useful in interpreting the data and guiding further analyses.

The results from the clustering analysis provided a deeper understanding of the data structure, revealing unique patterns and potential areas of interest. The clear delineation of data into specific clusters allowed for the identification of potential anomalies and marked the first step towards a comprehensive anomaly detection procedure. The insights gained from this analysis are essential for guiding subsequent investigations and enhancing the understanding of the data's inherent structure and patterns.

4.3.12. Anomaly Detection with Clustering

In the context of the conducted research, anomaly detection held a pivotal role, particularly when considered in tandem with the clustering analysis. This study employed the Isolation Forest algorithm for anomaly detection due to its notable competency in handling high-dimensional datasets. The algorithm was applied to the scaled data, previously organized into clusters, and the outputs were integrated into the original dataset as an "anomaly" column.

An intriguing aspect of the findings was the identification of 218 data points as potential anomalies amidst the 21,522 data points designated as non-anomalous. The "Count" variable played a crucial role in this differentiation, as the average "Count" for anomalies vastly exceeded that for non-anomalous data points. This substantial difference underscored the accuracy of the Isolation Forest algorithm in pinpointing outliers based on the "Count" variable.

A closer examination of the anomaly occurrence over time, as drawn from the summary analysis, revealed an insightful pattern. Anomalies began to appear from January 2017, initially at a modest rate of one per month. This frequency, however, gradually amplified, reaching an upsurge in September 2021 with 17 anomalies.

These observations provided invaluable insights into the behavior of the data over time and underlined the effectiveness of the anomaly detection method in combination with clustering analysis. The alignment of the areas of interest identified by both clustering and Isolation Forest algorithms enhanced the confidence in the robustness of the chosen methodologies.

The resulting findings offer a solid foundation for further investigation into these anomalies, which could lead to novel discoveries and enrich the existing body of knowledge in this domain. Future work is envisaged to explore these anomalies more intensively to elucidate their full implications, thereby contributing to a more profound comprehension of the data's structure and behavior.

4.4. Comparison to Previous Research

The results of the current study align with the existing body of research on network anomaly detection while also providing unique insights. In line with previous research, the

study reaffirms the role of machine learning in detecting network anomalies [18,21,25,30–32,35]. It further extends this by focusing on anomalies in network behavior characterized by unusual patterns potentially indicative of cyberthreats.

The behavioral scoring system used in this study is in line with the approaches by Alsarhan [30], Boateng [18], and Mengidis et al. [35], who also utilized machine learning methodologies for anomaly detection. The study, however, distinguishes itself by tying the scoring to a combination of Autonomous System Numbers and Names (ASNs), the country of origin, and the number of connections made, a more holistic approach to understanding network behavior.

The current research validates the significance of IP address and ASN in identifying anomalous network activities, an assertion also supported in the works of Alowaisheq [5] and Li [19]. It extends this understanding by providing quantifiable evidence through a behavior-scoring mechanism that links these factors with the frequency and nature of abnormal behavior, a contribution not previously articulated in such detail.

Echoing the works of Aboah Boateng [18] and Mengidis et al. [35], this study employs unsupervised machine learning methods for anomaly detection but differs in their application to a purely network-centric dataset. The current study also stresses the need for data reduction and dimensionality reduction techniques, a sentiment shared with the work of Fu et al. [34]. However, this study employs both source IP addresses and ASNs to carry out the reduction process, boosting the efficiency of anomaly detection.

The study concurs with Moriano Salazar's [44] emphasis on analyzing real-world temporal networks. It highlights the importance of continuous and real-time monitoring due to the dynamic nature of cyberthreats. Drawing on Alowaisheq's [5] work, it also examines network behavior from multiple angles, considering the origin of traffic and its associated behavior.

In line with Moriano Salazar [44] and Ongun's [25] discourse on the temporally dynamic nature of network behavior, the study further reinforces the need for constant model updates as cyberthreats evolve. It also integrates this concept into a practical framework, thereby offering actionable insights for the cybersecurity community.

The study diverges from previous research, focusing on a behavior-based scoring system linked to ASNs and the country of origin. While Chatterjee [26] used deep learning mechanisms for network intrusion detection, the current research offers a behavior-based scoring system as a potentially more accessible anomaly identification and severity assessment method.

The work of Christopher [45] underscores the significance of protecting the Industrial Control System (ICS) environment. The study builds upon this concept, emphasizing the role of ASNs and behavior scoring in improving defenses against intrusions. Similarly, the current research aligns with Wendt's work [7], reinforcing the strategies needed to enhance adaptive cyberdefenses, particularly in the financial sector. However, it further accentuates the role of behavior-based scoring and unsupervised learning in strengthening these defenses.

Aghaei's study [22] on the automated classification and mitigation of cybersecurity vulnerabilities resonates with the present research, which also emphasizes the automated detection of anomalies using machine learning techniques. The research diverges, however, in its application of these techniques specifically for network behavior analysis.

Research conducted by Bajic [14] underpins the importance of dynamic defense in computer networks. The current study complements this perspective, considering the dynamic nature of network behavior and the need for adaptive mechanisms to detect anomalies. The current research, however, extends this premise by developing a behavior-scoring system, contributing a novel approach to network anomaly detection.

The study aligns with the work of Villalón-Huerta and Ripoll-Ripoll [42] by emphasizing the importance of detecting and sharing behavioral indicators of compromise. It expands on this work by proposing a behavior-scoring system to classify and understand these indicators.

In summary, the present study builds upon and extends the knowledge in network anomaly detection, informed by prior research, while providing new insights through a unique behavior-based scoring system. These findings should be considered as a starting point for future research, refining and enhancing the understanding of network behavior anomalies.

4.5. Practical Implications and Recommendations

The findings of the longitudinal study have substantial practical implications for cybersecurity practitioners, particularly those working in network security. Understanding the impact of these results can guide the formulation of effective anomaly detection and mitigation strategies and eventually help enhance overall network security.

1. The analysis points to the need for a comprehensive and multi-pronged approach to network security. Practitioners should integrate temporal, geographical, and autonomous system data along with a quantitative measure of behavior [40]. This multi-faceted approach will allow practitioners to detect, understand, and respond to increasingly sophisticated and targeted attacks more effectively.
2. Practitioners need to focus on developing and deploying specialized preventative and responsive measures for each unique network characteristic. The characteristic of network traffic data should guide these measures [40]. The cluster will require continuous monitoring and updating of security measures to match the evolving nature of cyberthreats.
3. The time-series analysis suggests a requirement for security systems that can effectively identify periods of unusual activity or anomalies. Given the rapidly evolving nature of cyberattack techniques, this temporal approach to security is becoming increasingly important [43].
4. As a quantitative measure of potential anomalies, the behavior score should be utilized in security systems to detect abnormal behavior. The behavior score can be vital in gauging the increasing sophistication and target specificity of cyberattacks [39].
5. The high frequency of abnormalities originating from specific geographical locations and autonomous systems suggests a need for closer monitoring and stricter security measures in these areas [38].
6. Practitioners must remember that while the behavior score effectively detects abnormalities, it does not pinpoint the exact type or severity of the anomaly [39]. Therefore, future research and development efforts should refine this tool to provide more detailed information about detected anomalies. This tool would help design more precise and effective responses [41].

Considering the implications above, it is recommended that organizations adapt their network security strategies to integrate the insights gained from this study. These strategies would involve updating monitoring practices to include clustering and time-series analysis, employing the behavior scoring system, and paying particular attention to high-risk geographical locations and autonomous systems.

While the longitudinal study has provided valuable insights, network security is constantly evolving. As new patterns of network behavior emerge and new types of threats are devised, continuous research and development in network security are essential to stay ahead of potential security risks. The methodologies employed in this study can serve as a foundation for future research in this critical area of cybersecurity.

4.6. Limitations and Future Research

While the longitudinal study has generated substantial insights into network anomaly detection and associated implications, it is essential to recognize its limitations, which can also serve as potential directions for future research.

1. One limitation of the study is the inherent restriction in the behavior score, which quantifies the likelihood of network anomalies. However, it falls short of identifying the exact type or severity of the anomaly [39]. While it proved effective in recogniz-

ing unusual behavior patterns, enhancing this tool to provide precise details about anomalies detected will be crucial in future research endeavors.

2. Another drawback is related to the dependency on certain types of data-time-based, location-specific, and autonomous system data. Although these data types are crucial in network security, examining other data categories could foster a deeper, more multifaceted comprehension of network traffic behavior and security threats [43].
3. The geographical analysis revealed a higher frequency of abnormalities from specific locations and autonomous systems [38]. However, given the global and interconnected nature of cyberthreats, future research could benefit from a more expansive geographical scope that includes a wider variety of regions and autonomous systems.
4. The research retrospectively identified abnormalities and anomalies using past data [40]. Consequently, the study suggests a need for more proactive security measures. Therefore, future research might explore developing and testing predictive models, enabling practitioners to anticipate and counter cyberthreats before they occur.
5. With a solely quantitative approach, the current study relies predominantly on numerical data. Future research could benefit from incorporating qualitative methods into the analysis, such as expert opinions or case studies. Including qualitative methods would enable a more comprehensive understanding of network anomalies and their associated threats, thereby expanding the scope of the study's results.
6. There were certain overlooked factors in this study. It did not consider variables like the type of network protocol, the application linked to the network traffic, and specific details about the source and target systems. Future research that includes these overlooked factors could enrich the analysis, providing additional dimensions that lead to a more detailed and robust understanding of network behavior.
7. Finally, the research was limited to using specific techniques for data analysis. While these techniques provided valuable insights, future research could explore the integration of emerging methodologies such as deep learning techniques [43], behavior-based threat-hunting frameworks [39], and advances in intrusion detection techniques [41]. These emerging methodologies could provide a richer and more in-depth understanding of network traffic behavior and cyberthreats, contributing to advancing network security practices.

The limitations of the current study offer avenues for future research. The dynamic nature of network behavior and cyberthreats necessitates ongoing research in this area. Future studies should continue to evolve and expand on the methodologies used in this study, incorporating more comprehensive data and refining the analysis techniques. Doing so will enhance our understanding of network anomalies and their threats, improving our capacity to safeguard our networks against cyberthreats.

5. Conclusions

5.1. Summary of Main Findings

The collective findings from the longitudinal research provide considerable advancements in understanding network traffic behavior's diversity and complexity, affirming the study's initial research question and hypothesis. This understanding revolves significantly around detecting network anomalies, emphasizing the integration of geographical, organizational, and behavioral analyses.

1. A pivotal insight revealed from the research shows that cyberattacks have grown increasingly sophisticated and targeted. This growth was demonstrated by the time-series analysis and clustering of network traffic data, with specific periods of abnormal activity being identified [38]. The time-series analysis and clustering underscore the importance of temporal awareness in network security endeavors.
2. The study also confirmed the efficacy of a behavior score ranging from 0 to 136 in measuring network anomalies. This scoring system reflects cyberattacks increased complexity and target specificity [39]. However, it is crucial to acknowledge this system's limitation in precisely defining the anomaly's type or severity.

3. Geographical data and Autonomous System Numbers (ASNs) were found to play a significant role in decoding network anomalies [38]. Higher frequencies of irregularities from specific regions and systems indicated the need for heightened scrutiny and security measures.
4. Additionally, through organizational analysis, specific organizations, including "DigitalOcean-ASN", "F3 Netze e.V.", and "Zwiebelfreunde e.V.", were frequently associated with IP addresses possessing high behavior scores. This observation underlines the importance of considering ASNs and organizational information in detecting network anomalies.
5. This study further underscores the multifaceted nature of network traffic behavior, emphasizing the necessity for comprehensive, multi-pronged approaches to network security [40]. The dynamic nature of cyberthreats calls for incorporating various data types—temporal, geographical, and autonomous system data, coupled with a quantitative measure of behavior—for effective identification and mitigation of network anomalies in a changing threat landscape.
6. Moreover, the robustness of the findings was ensured through the application of cross-validation techniques. This cross-validation emphasized the need for continual updates and recalibrations of the models in line with the evolution of cyberthreats.

In conclusion, the present study highlights the changing nature of network security threats, underscoring the demand for increasingly sophisticated and comprehensive security approaches. The findings align with previous research and provide valuable insights for future investigations in the field, thus informing the development of more nuanced and context-aware anomaly detection systems. This study, therefore, offers practical implications for both academia and industry.

5.2. Contributions to the Field

The longitudinal research holds significant implications for the field of network security, primarily through its comprehensive examination of network traffic behavior and its consequential implications for addressing the evolving landscape of cyberthreats. The findings also contribute to enriching the understanding and methodological approaches to network anomaly detection, a critical area in cybersecurity.

1. One of the crucial contributions of this study is the development and validation of a behavior scoring system as an effective tool to gauge the sophistication and target specificity of cyberattacks [39]. This approach offers a novel, quantifiable measure of potential anomalies in network behavior.
2. By grouping network traffic data into distinct clusters, the research provides an innovative method to uncover and understand the diversity in network behavior patterns [38]. The clusters and behavior patterns add to our understanding of network traffic behavior and aid in designing targeted preventative and responsive measures.
3. The application of time-series analysis and geographical and autonomous system data has demonstrated the potential of a comprehensive approach to network security, which integrates multiple data types to identify and mitigate network anomalies [40] effectively.
4. By pinpointing periods of abnormal activity, the research has contributed to the field's understanding of temporal patterns in cyberattacks, highlighting the necessity for temporal considerations in network security efforts.
5. By identifying regions and systems with higher frequencies of network anomalies, the research has underscored the importance of geography and autonomous system data in cybersecurity efforts.
6. The empirical evidence provided in the study establishes transparent relationships between geographical, organizational, and behavioral aspects and network anomalies. This practical grounding enriches the theoretical underpinnings of the field and extends real-world applicability to the findings.

In summary, this study has provided an integrative and multifaceted approach to understanding network traffic behavior and cyberthreat detection. It contributes valuable insights to the field and offers practical tools and methods that can be used in future research and the development of robust network security measures.

5.3. Practical Implications

The outcomes of the longitudinal research hold substantial practical implications that can directly benefit cybersecurity professionals, organizations, and network administrators. These implications manifest across multiple dimensions of network security, from enhanced network monitoring to cybersecurity training and education.

1. The behavior scoring system developed in this research provides a pragmatic tool for detecting potential anomalies in network traffic. This system contributes a quantifiable measure of cyberattack sophistication and target specificity, which can be used in proactive threat detection efforts [39]. Simultaneously, it allows for a quantitative assessment of network anomalies, proving invaluable for risk management.
2. The research introduces an integrative approach to network anomaly detection by integrating geographic, organizational, and behavioral factors. This unified model yields a more comprehensive understanding of network behavior that can inform real-time network monitoring and improve the identification and response to potential threats.
3. As highlighted by this study, recognizing the role of geographic and organizational context in network anomalies equips organizations to develop tailored cybersecurity strategies. For instance, stricter controls or more rigorous monitoring could be applied for network traffic from countries or organizations associated with higher behavior scores.
4. The categorization of network traffic data into distinct clusters offers network security professionals an innovative methodology to comprehend the diversity in network behavior patterns [38]. This approach can assist in designing more targeted preventive measures.
5. Additionally, understanding temporal patterns in cyberattacks carries implications for the timing of security measures. As time-series analysis indicates, organizations may need to bolster their defenses during periods of unusual activity.
6. Findings related to the sources of network anomalies suggest that specific geographical regions and autonomous systems warrant scrutiny. This insight underlines the need for more tailored strategies to ensure network security in these areas.
7. Moreover, systems such as "DigitalOcean-ASN", "F3 Netze e.V.", and "Zwiebelfreunde e.V.", identified as frequent sources of abnormalities, indicate that companies and organizations using these systems should consider implementing additional security measures.
8. The validation of the scoring mechanism provides a standard measure for network anomalies that can be used for benchmarking and predictive analysis. This benchmark allows for comparisons of network behaviors across time and context, aiding organizations to anticipate and respond proactively to potential threats.
9. The insights derived from this study hold potential for integration into cybersecurity training programs. The behavior scoring mechanism can be an effective teaching tool, helping practitioners understand the multi-faceted contributors to abnormal network behaviors.

In conclusion, the practical implications of this longitudinal research carry the potential to enhance network security practices significantly. The implications provide valuable insights into network traffic behavior and offer innovative tools for proactive cyberthreat detection and response, strengthening defenses against an evolving threat landscape.

5.4. Regarding Future Research Directions

Future research could enhance the understanding of cyberthreats and refine strategies to mitigate them by building on the findings of this study, with the following directions being considered crucial:

1. Expanding Understanding of Anomaly Behavior: A deeper exploration into the nature and implications of detected anomalies could be a focus for future studies. By refining the behavior score system, a more comprehensive understanding of the types and severity of anomalies could be achieved [39].
2. Examining Network Behavior Clusters: Future research could delve into the specifics of the identified network behavior clusters. A more detailed understanding could guide the creation of targeted security measures and accommodate the diversity in network behavior patterns [38].
3. Enhancing Temporal Analysis Techniques: Leveraging machine learning and artificial intelligence to predict periods of heightened cyberattack activity could enhance the proactive capabilities of security systems. Research that predicts these patterns over an extended period could yield beneficial insights.
4. Investigating Geographic and Systemic Anomalies: Understanding why certain regions and systems are more prone to network anomalies would be beneficial. These insights could enable the development of preventative measures specific to certain regions or autonomous systems.
5. Exploring Alternatives in Data Handling Platforms: Testing the utility of various data handling platforms, in line with the transfer of the "UWF-ZeekData22" to Hadoop's distributed file system, could offer comparative insights on the efficiency and reliability of different systems for network security research [38].
6. Utilizing AI and Machine Learning: Exploring the potential of advanced techniques like deep learning in network anomaly detection could significantly augment the capacity to handle complex and evolving threats [43].
7. Focusing on IoT-Specific Threats: Given the growing prevalence of IoT devices, dedicated studies into IoT malware behavior and intrusion detection techniques are becoming increasingly necessary [41].
8. Examining Identified Autonomous Systems: Studies targeting the security vulnerabilities of systems identified as frequent sources of abnormalities, such as "DigitalOcean-ASN", "F3 Netze e.V.", and "Zwiebelfreunde e.V". could help develop system-specific security measures.
9. Investigating Human Factor and Regulatory Frameworks: The role of the human factor in cybersecurity and understanding global regulatory and policy frameworks for anomaly detection is another promising research direction. Future research could explore improving user awareness and behavior and how anomaly detection impacts regulatory policies.
10. Privacy-Preserving Anomaly Detection: Considering growing privacy concerns, future investigations could explore developing anomaly detection techniques that respect user privacy, such as differential privacy or federated learning.

These potential research directions could enrich our understanding of the dynamic cyberthreat landscape and contribute to developing more robust network security strategies in an increasingly interconnected world.

5.5. Final Thoughts

The current era of accelerated technological advancement has witnessed an unprecedented escalation in the complexity and sophistication of cyberthreats. This longitudinal study serves as a vital stepping stone in understanding and mitigating these threats by revealing patterns in network traffic behavior, offering a statistical view of the fluctuations in attack counts, and emphasizing the necessity of a comprehensive approach to network security.

The results presented herein demonstrate meaningful strides in cyberthreat detection, reinforcing the continuous battle against evolving cyberattacks. As these threats become increasingly targeted, the understanding and defensive mechanisms must adapt in tandem. Incorporating temporal, geographical, and autonomous system data into security strategies is underscored as a suggestion and a necessity in navigating the fluid threat landscape.

Moreover, the significance of AS-based anomaly detection in combatting evolving cyberthreats is brought to the fore by this research. An essential lesson is the need for constant evolution and enhancement of anomaly detection systems to keep pace with the increasingly sophisticated digital threats of our time. This demanding task, while formidable, is a pursuit that must be unwavering, given the growing reliance on digital infrastructures.

Integrating diverse data points, such as source IP addresses, AS numbers, and AS organization names, is an effective strategy for enhancing anomaly detection accuracy and efficiency. This approach harnesses the strength of data, a pivotal asset in today's digital era, furnishing us with better defenses against potential cyberthreats.

However, it is acknowledged that combat against cyberthreats is multifaceted, necessitating a dynamic strategy and constant updates to cybersecurity measures reflecting the rapidly evolving threat landscape. The pivotal role of collaboration between stakeholders, at both technical and policy levels, is emphasized, with a call to cultivate a culture of information sharing and cooperative action against common threats.

Furthermore, while providing valuable insights, this study also underscores future research requirements. This work aims to employ emerging technologies, such as artificial intelligence, machine learning, and quantum computing, to boost our cybersecurity capabilities. Simultaneously, it is imperative to scrutinize the implications of these technologies on privacy and regulatory norms.

As the Internet continues to embed itself in society's fabric, the importance of robust cybersecurity measures, including efficient anomaly detection, will inevitably rise. Pursuing a secure digital world is ongoing, with this research aspiring to contribute positively to the collective endeavor. This journey calls for the united efforts of researchers, practitioners, policymakers, and users alike, aiming to create an environment where security measures match and surpass the sophistication of the threats faced, steering us toward a more secure digital future.

Funding: This research received no external funding. The author incurred all costs.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Access to the study data and associated code can be granted upon request to the corresponding author. Public availability is not provided in order to maintain controlled access and to protect the integrity of both the data and the code.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Farokhnia Hamedani, M. *Essays on Cybersecurity and Information Privacy*; ProQuest Dissertations Publishing, University of South Florida: Tampa, FL, USA, 2023.
2. Rosa, F.R. *Global Internet Interconnection Infrastructure: Materiality, Concealment, and Surveillance in Contemporary Communication*; ProQuest Dissertations Publishing, American University: Washington, DC, USA, 2019.
3. Weathersby, A. Discerning the Relative Threat of Different Network Based Cyber-Attacks, a Study of Motivation, Attribution, and Anonymity of Hackers. Ph.D. Thesis, Marymount University, Arlington, WV, USA, 2023.
4. Myneni, S. Defeating Attackers by Bridging the Gaps Between Security and Intelligence. Ph.D. Thesis, Arizona State University, Arizona, WV, USA, 2022.
5. Alowaisheq, E. *Security Traffic Analysis Through the Lenses of: Defenders, Attackers, and Bystanders*; ProQuest Dissertations Publishing, Indiana University: Bloomington, IN, USA, 2020.
6. Barron, T. *Addressing the Imbalance between Attackers and Defenders Using Cyber Deception*; ProQuest Dissertations Publishing, State University of New York at Stony Brook: Stony Brook, NY, USA, 2020.
7. Wendt, D.W. Exploring the Strategies Cybersecurity Specialists Need to Improve Adaptive Cyber Defenses within the Financial Sector: An Exploratory Study. D.C.S. Dissertation, Colorado Technical University, Colorado Springs, CO, USA, 2020.
8. Adewopo, V. *Exploring Open Source Intelligence for Cyber Threat Prediction*; ProQuest Dissertations Publishing, University of Cincinnati: Cincinnati, OH, USA, 2021.
9. Cho, S. *Tackling Network-Level Adversaries Using Models and Empirical Observations*; ProQuest Dissertations Publishing, State University of New York at Stony Brook: Stony Brook, NY, USA, 2021.

10. Muoi, T.D. *Handling Network Attacks Exploiting Routing Information Asymmetries*; ProQuest Dissertations Publishing, National University of Singapore: Singapore, 2022.
11. Panagiotou, P.; Mengidis, N.; Tsikrika, T.; Vrochidis, S.; Kompatsiaris, I. *An in Depth Analysis of Open Source Tools: Host Intrusion Detection System, Intrusion Detection System, and Honeypots, and How They Can Protect a SME's Network*; ProQuest Dissertations Publishing, Utica College: Utica, NY, USA, 2019.
12. Andrews, K.T. Deception Techniques and Technologies in the Role of Active Cyber Defense. Master's Thesis, Utica College, Utica, NY, USA, 2020.
13. Gutierrez, M. Detecting Complex Cyber Attacks Using Decoys with Online Reinforcement Learning. Ph.D. Thesis, The University of Texas at El Paso, El Paso, TX, USA, 2023.
14. Bajic, A. Simulation-Based Evaluation of Dynamic Attack and Defense in Computer Networks. Ph.D. Thesis, Freie Universitaet Berlin, Berlin, Germany, 2021.
15. Bobish, M. *Sharing Cyber Threat Information Between the United States' Public and Private Sectors*; ProQuest Dissertations Publishing, Utica University: Utica, NY, USA, 2023.
16. Abu, M.S.; Selamat, S.R.; Yusof, R.; Ariffin, A. Formulation of Association Rule Mining (ARM) for an Effective Cyber Attack Attribution in Cyber Threat Intelligence (CTI). *Int. J. Adv. Comput. Sci. Appl.* **2021**, *12*, 4. [CrossRef]
17. Parker, C.M. Exploring the Use of Information Security Practices in Response to Cyberattacks to Protect U.S. Federal Systems and Networks. Ph.D. Thesis, Northcentral University, San Diego, CA, USA, 2021.
18. Aboah Boateng, E. *Unsupervised Machine Learning Methods for Detecting Process Control Anomalies in Industrial Control Systems*; ProQuest Dissertations Publishing, Tennessee Technological University: Cookeville, TN, USA, 2023.
19. Li, G. *An Empirical Analysis on Threat Intelligence: Data Characteristics and Real-World Uses*; ProQuest Dissertations Publishing, University of California: San Diego, CA, USA, 2020.
20. Gyamfi, E.; Jurcut, A. Intrusion Detection in Internet of Things Systems: A Review on Design Approaches Leveraging Multi-Access Edge Computing, Machine Learning, and Datasets. *Sensors* **2022**, *22*, 3744. [CrossRef] [PubMed]
21. Shin, Y.; Kim, K. Comparison of Anomaly Detection Accuracy of Host-based Intrusion Detection Systems based on Different Machine Learning Algorithms. *Int. J. Adv. Comput. Sci. Appl.* **2020**, *11*, 252–259. [CrossRef]
22. Aghaei, E. Automated Classification and Mitigation of Cybersecurity Vulnerabilities. Ph.D. Thesis, The University of North Carolina at Charlotte, Charlotte, NC, USA, 2022.
23. Hillis, J.S. *Enterprise Advanced Persistent Threat Group Identification and Technique Discovery*; ProQuest Dissertations Publishing, Marymount University: Arlington, WV, USA, 2023.
24. Luitel, A. *A Framework for Modeling Data Breach Risk Using Machine Learning Models for High-Dimensional Panel Data*; ProQuest Dissertations Publishing, The George Washington University: Washington, DC, USA, 2022.
25. Ongun, T. *Resilient Machine Learning Methods for Cyber-Attack Detection*; ProQuest Dissertations Publishing, Northeastern University: Boston, MA, USA, 2023.
26. Chatterjee, S. *Network Intrusion Detection and Deep Learning Mechanisms*; ProQuest Dissertations Publishing, Florida Atlantic University: Boca Raton, FL, USA, 2023.
27. Masarweh, A.A. Enhancing the Penetration Testing Approach and Detecting Advanced Persistent Threat Using Machine Learning. Master's Thesis, Princess Sumaya University for Technology, Amman, Jordan, 2021.
28. Rahman, A.; Ali, A.; Iqbal, F.; Hussain, M.; Ullah, F. Deep Learning Methods for Malware and Intrusion Detection: A Systematic Literature Review. *Secur. Commun. Netw.* **2022**, *2022*, 2959222. [CrossRef]
29. Villanueva-Miranda, I. Modeling and Predicting Emerging Threats Using Disparate Data. Ph.D. Thesis, The University of Texas at El Paso, El Paso, TX, USA, 2023.
30. Alsarhan, H.F. *Real-Time Machine Learning-based Intrusion Detection System (IDS) for Internet of Things (IoT) Networks*; ProQuest Dissertations Publishing, The George Washington University: Washington, DC, USA, 2023.
31. Al-Haija, Q.A.; Krichen, M.; Elhaija, W.A. Machine-Learning-Based Darknet Traffic Detection System for IoT Applications. *Electronics* **2022**, *11*, 556. [CrossRef]
32. Butt, S.M.; Reaiche, C. Cognitive Analysis of Intrusion Detection System. *J. Sib. Fed. University. Eng. Technol.* **2022**, *15*, 102–120. [CrossRef]
33. Mahfouz, A.M.; Abuhussein, A.; Alsubaei, F.S.; Shiva, S.G. Toward A Holistic, Efficient, Stacking Ensemble Intrusion Detection System using a Real Cloud-based Dataset. *Int. J. Adv. Comput. Sci. Appl.* **2022**, *13*, 950–962.
34. Fu, X.; Zhang, Y.; Li, H.; Hu, Y. Research on Attributes Reduction Method of Intrusion Detection Data Based on Rough Set Theory. *J. Phys. Conf. Ser.* **2020**, *1624*, 032036. [CrossRef]
35. Mengidis, N.; Panagiotou, P.; Tsikrika, T.; Vrochidis, S.; Kompatsiaris, I. Host-based Intrusion Detection Using Signature-based and AI-driven Anomaly Detection Methods. *Inf. Secur.* **2021**, *50*, 37–48. [CrossRef]
36. Moore, K.E. *Analyzing Small Business Strategies to Prevent External Cybersecurity Threats*; ProQuest Dissertations Publishing, Walden University: Minneapolis, MN, USA, 2023.
37. Phillips, I.J., Jr. *Maintaining Small Retail Business Profitability by Reducing Cyberattacks*; ProQuest Dissertations Publishing, Walden University: Minneapolis, MN, USA, 2020.
38. Bagui, S.S.; Mink, D.; Bagui, S.C.; Ghosh, T.; Plenkers, R.; McElroy, T.; Dulaney, S.; Shabanali, S. Introducing UWF-ZeekData22: A Comprehensive Network Traffic Dataset Based on the MITRE ATT & CK Framework. *Data* **2023**, *8*, 18. [CrossRef]

39. Bhardwaj, A.; Kaushik, K.; Alomari, A.; Alsirhani, A.; Alshahrani, M.M.; Bharany, S. BTH: Behavior-Based Structured Threat Hunting Framework to Analyze and Detect Advanced Adversaries. *Electronics* **2022**, *11*, 2992. [CrossRef]
40. Ntingi, N.; Duvenage, P.; du Toit, J.; von Solms, S. Effective Cyber Threat Hunting: Where and how does it fit? In Proceedings of the European Conference on Cyber Warfare and Security, Reading, UK, 16–17 June 2022; pp. 206–213.
41. Davanian, A. Techniques for Detecting Intrusions. Ph.D. Thesis, University of California, Riverside, CA, USA, 2022.
42. Villalón-Huerta, A.; Ripoll-Ripoll, I.; Marco-Gisbert, H. Key Requirements for the Detection and Sharing of Behavioral Indicators of Compromise. *Electronics* **2022**, *11*, 416. [CrossRef]
43. Torre, D.; Mesadieu, F.; Chennamaneni, A. Deep Learning Techniques to Detect Cybersecurity Attacks: A Systematic Mapping Study. *Empir. Softw. Eng.* **2023**, *28*, 76. [CrossRef]
44. Moriano Salazar, P. *Anomaly Detection in Real-World Temporal Networks*; ProQuest Dissertations Publishing, Indiana University: Bloomington, IN, USA, 2019.
45. Christopher, C.I. Protecting the Industrial Control System Environment: Implementing Active Cyber Defense to Aid Mitigation of Threat Intrusions. Master's Thesis, Utica College, New York, NY, USA, 2020.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.