The Pennsylvania State University
The J. Jeffrey And Anne Marie Fox Graduate School

LEVERAGING LARGE DARKNETS FOR ACTIONABLE THREAT INTELLIGENCE: AN ARTIFICIAL INTELLIGENCE-DRIVEN APPROACH

A Dissertation in
Informatics
by
Rupesh Prajapati

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

December 2024

The dissertation of Rupesh Prajapati was reviewed and approved by the following:

Dinghao Wu
Professor of College of IST
Dissertation Advisor
Chair of Committee

John Yen
Professor of College of IST
Committee Member

Vasant Honavar
Professor of College of IST
Committee Member

Jia Li
Professor of Eberly College of Science
Committee Member

Michalis Kallitsis
Akamai Technologies, Inc.
Special Member

Carleen Maitland
Program Head

Abstract

Adversaries increasingly rely on active reconnaissance techniques, such as probing, to identify and exploit vulnerabilities within target systems. Understanding these probing activities provides invaluable insights into the evolving threat landscape, empowering security professionals to proactively adapt their defense strategies and mitigate potential cyberattacks. Large network telescopes, or darknets, offer a powerful resource for analyzing these probes in detail, capturing vast amounts of scanning traffic from a wide range of potential malicious actors. However, effectively extracting timely, actionable threat intelligence from this massive volume of darknet data remains a significant challenge. This dissertation explores the potential of artificial intelligence (AI) techniques to overcome these challenges and provide actionable threat intelligence from large-scale darknet data.

Focusing on the intricate interplay of scanning behavior, system vulnerabilities, and evolving threat actor tactics, this research investigates the efficacy of AI for extracting actionable intelligence from darknet data. First, the study delves into the identification and interpretation of temporal changes within the Internet threat landscape. By analyzing network traffic patterns and identifying anomalies in scanning behaviors, this research presents a novel framework for near real-time detection of emerging threats. Furthermore, the dissertation explores the potential for cross-sensor data fusion, leveraging the insights gleaned from darknets to correlate with intelligence gathered from other security sensor networks. This enables the inference of threat actor motives and techniques based solely on their scanning behavior, further enhancing the granularity of threat intelligence. Lastly, this research investigates the feasibility of a Learning Using Privileged Information (LUPI) framework to significantly enhance threat intelligence inference. By incorporating limited but highly valuable privileged information, this approach seeks to improve the accuracy and efficiency of AI-driven threat intelligence extraction from darknet data. This dissertation ultimately contributes to a more comprehensive understanding of the cyber threat landscape, enabling the development of robust and proactive security strategies.

Table of Contents

List of Figures

List of Tables

Acknowledgments

This dissertation would not have been possible without the support of many individuals.

First, I wish to express my deepest gratitude to my adviser, Dr. Dinghao Wu, for his unwavering support, guidance, and invaluable advice throughout my research journey. His patience and understanding, especially during my periods of uncertainty regarding my dissertation topic, have been immensely appreciated.

I would also like to extend my heartfelt thanks to my esteemed committee members—Dr. John Yen, Dr. Vasant Honavar, Dr. Jia Li, and Dr. Michalis Kallitsis[1]—for their insightful suggestions and constructive feedback, which have significantly enhanced the quality of this work. I am particularly grateful to Dr. Yen for his constant availability and detailed input on all aspects of my research, as well as for his thorough reviews of numerous drafts. My interactions with Dr. Honavar have been invaluable, providing clarity and direction during challenging phases of my research. I appreciate Dr. Li for her willingness to join my committee on short notice, which contributed greatly to the depth of my study. I would also like to extend a special thanks to Dr. Kallitsis; this research took flight thanks to his continuous support and guidance.

My journey to this point would not have been possible without the incredible encouragement, sacrifices, and love from my parents, Bishnu Laxmi Prajapati and Laxmi Prasad Prajapati. I am also grateful to my brother, Rujin Prajapati, for his unwavering support and encouragement. The friendships I've built in State College have made my time here unforgettable. In particular, I would like to thank Dr. Bikalpa Neupane, Dr. Bipin Rimal, Dr. Santosh Panthi, and Dr. Prasanna Umar; their mentorship and friendship have been invaluable to my personal and professional growth.

Lastly, I am profoundly grateful to Merit Network, Inc. for providing the computational resources and data essential for this research. Also, I would like to thank GreyNoise Intelligence, Inc. for supporting this research by providing threat intelligence data.

This dissertation was partially funded by the U.S. Department of Homeland Security under Grant Award Number 17STQAC00001-05-00, and the National Science Foundation under awards CNS-1823192, CNS-2120400 and CNS-2213794. The opinions, findings, and conclusions expressed herein are those of the author and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Government.

[1]Collaboration with Dr. Kallitsis took place during his tenure with Merit Network, Inc., University of Michigan.

Chapter 1
Introduction

The explosive growth of the Internet has fostered a vastly interconnected yet vulnerable ecosystem, characterized by widespread adoption of inadequate security practices and prevalent use of easily guessable default credentials and outdated software, creating a fertile ground for cyberattacks. This vulnerability is compounded by the continuous refinement of sophisticated adversarial techniques and the readily available tools for automated vulnerability scanning, such as ZMap [2] and Masscan [3], capable of rapidly scanning the entire IPv4 address space. The consequences of such attacks, as evidenced by high-profile incidents such as the Mirai botnet distributed denial-of-service (DDoS) attack against Dyn in 2016 [4], the Equifax data breach in 2017 [5], and the NotPetya and WannaCry ransomware outbreaks in the same year [6, 7], demonstrate the significant financial and reputational damage inflicted on individuals and organizations alike.

These attacks frequently initiate with a reconnaissance, a crucial phase in the cyber kill chain [8], employing techniques such as port scanning and network mapping to identify vulnerable systems. The proliferation of high-speed scanning tools significantly accelerates this phase, enabling large-scale probing campaigns and facilitating rapid exploitation by bot herders [9] and self-propagating malware [10]. The established correlation between such prior scanning activities and subsequent malware infection [11] has been exploited for proactive threat detection.

The use of strategically deployed sensor networks such as network telescopes provide a unique vantage point for monitoring such malicious activities. Network telescopes, also commonly referred to as darknets, passively monitor and collect traffic directed at unassigned IP addresses ("dark IP space"). The unsolicited traffic captured by these darknets constitutes a significant portion of Internet Background Radiation (IBR), which predominantly consists of unwanted traffic generated by botnets and scanning tools [12, 13]. This dissertation investigates the application of artificial intelligence (AI),

machine learning and other statistical tools to enhance the ability to derive actionable threat intelligence from the darknet traffic data.

This unsolicited nature of the captured traffic, owing to the unadvertised nature of the darknet IPs, significantly increases the probability that observed communication is malicious, thus transforming darknet data into a valuable resource for analyzing adversarial behavior. The scalability inherent in network telescope deployments, capable of monitoring hundreds of thousands of IP addresses, enables the acquisition of comprehensive and high-resolution data that captures the broad scope of malicious network activity. However, the substantial volume and specific characteristics of this data necessitate sophisticated machine learning models for efficient analysis.

The central focus of this research is to elucidate the challenges inherent in fully leveraging darknet data and to address these challenges through the development of robust methodologies that can address the high-dimensional feature space. This involves formulating the complexities of the problem into solvable frameworks, followed by the rigorous development and evaluation of predictive models. In this process, it is imperative to acknowledge various influencing factors—both manifest and latent—such as the continuously evolving nature of malicious patterns, the challenges posed by imbalanced class distributions, and the intricate correlations that may exist within the data.

In the field of cybersecurity, the analysis of darknet data has yielded critical insights into a variety of network anomalies, including, but not limited to, port scanning activities [14–16], worm propagation dynamics [17–21], denial-of-service (DoS) attacks, as well as incidents stemming from misconfigurations [22], internet outages [23, 24], and politically motivated cyber operations. These analyses have significantly contributed to the identification and mitigation of malicious activities, proving especially valuable in post-breach investigations by facilitating attack reconstruction and the tracing of their origins. However, the potential of utilizing darknet data for near real-time threat intelligence remains relatively underexplored.

The scale and deployment of network telescopes directly influence data collection capabilities. Large darknets provide enhanced resolution and comprehensive coverage, resulting in massive datasets that facilitate a more nuanced analysis of internet traffic patterns. This increased resolution enables precise identification of traffic sources, destinations, and types, while the increased data points enhance the representativeness of the sample, thereby improving analytical accuracy and reliability. Notably, larger darknets exhibit heightened sensitivity to low-level, low-frequency traffic, which is often

indicative of stealthy, protracted attacks, a capability frequently lacking in smaller systems. Additionally, long-term monitoring capabilities further contribute to the identification of these slow and evasive attack vectors. This research leverages data from one of the largest network telescopes, ORION (Observatory for Cyber-Risk Insights and Outages of Networks) Network Telescope [1], which monitors nearly half a million IP addresses and captures more than 2GB traffic data daily.

The extensive dataset recorded by the ORION darknet encompasses a rich variety of information concerning incoming traffic, leading to a high-dimensional feature space. This high dimensionality presents considerable challenges in the analysis of the data. A substantial portion of this research is dedicated to developing a universal low-dimensional representation of such high-dimensional data, which is essential for subsequent analytical processes. This aspect becomes particularly critical when employing machine learning models, as their efficacy is heavily contingent upon the representation of the data. Effective data representation is instrumental in facilitating the identification of meaningful patterns within the inherent complexities of the dataset.

The initial phase of cybersecurity research invariably centers on the detection of malicious activities, with subsequent analysis and inference following this critical step. When confronted with large volumes of data, the instinctual approach often involves clustering to discern emergent groups. While this technique may successfully identify clusters of malicious activities, it falls short in elucidating the temporal evolution of these groups—a factor that is particularly crucial in the context of malware analysis. Understanding how these clusters evolve over time is essential for developing effective strategies for threat detection and mitigation. The first research question that this dissertation aims to answer is: How can the scanning activities be effectively profiled and analyzed to detect and interpret the temporal evolution of malicious activities in the Internet threat landscape, thereby enhancing the understanding of malicious behavior and informing proactive defense strategies? This inquiry aims to connect static clustering methods of darknet data with a deeper understanding of the dynamic nature of cyber threats. The goal is to enhance real-time threat intelligence capabilities.

The passive nature of network telescopes implies that they neither initiate connections nor respond to incoming requests; consequently, they exclusively collect unidirectional traffic without completing TCP handshakes. This design choice is intentional, as attempting to capture and store all data from such interactions would be logistically unfeasible. However, this limitation presents significant challenges for forensic analysis utilizing darknet data. In contrast, smaller sensor networks like honeypots actively engage with

malicious actors, allowing them to capture a dynamic threat profile. The second research question this dissertation seeks to address is whether the broader observability afforded by network telescopes can be synergistically integrated with threat intelligence gathered from smaller networks, thereby facilitating a more comprehensive and expedited understanding of emerging threats.

In the conventional machine learning paradigm, a feature vector is paired with a corresponding output label, enabling the model to learn the mapping from features to outcomes. In the domain of cybersecurity, a wealth of information is derived from both static and dynamic analyses. However, this information often remains inaccessible during inference, rendering it unsuitable as features for model training and typically leading to its exclusion from the modeling process. The Learning Using Privileged Information (LUPI) paradigm addresses this limitation by allowing these auxiliary insights to serve as supervisory signals during model training. This approach facilitates the model's acquisition of more nuanced and robust associations between the feature vector and the output labels. Specifically, the insights gleaned from smaller sensors, which distill threat intelligence labels, can be leveraged as privileged information to enhance the model's learning capacity. This, in turn, improves performance during testing phases, even when such privileged information is not available. The third research question of this dissertation investigates the feasibility of utilizing privileged information to improve performance of models aimed at analyzing and interpreting darknet scanning traffic.

Throughout the endeavor to address these questions, this dissertation confronts various challenges and issues, including data correlation and the absence of ground truth labels prevalent in cybersecurity research. This work contributes to the advancement of the field by presenting a robust methodology for effectively leveraging the rich, albeit complex, information contained within the darknet.

1.1 Scope

While the IBR collected by ORION encompasses a variety of traffic types, including backscatter traffic and traffic resulting from misconfigurations, this research is exclusively concentrated on scanning activities. This focus is justified, as scanning traffic constitutes over 90% of the overall dataset, making it a significant area of interest for threat detection.

Furthermore, it should be noted that the tools and techniques developed in this research were primarily designed with large darknets in mind. Consequently, their effectiveness may be diminished when applied to small or medium-scale darknets. The

methodologies demonstrated herein have proven successful within the contexts of extensive darknets and widely adopted honeypot implementations. However, adjustments may be necessary to adapt these techniques for application to less common sensor networks.

This acknowledgment of the limitations in scalability and adaptability underscores the importance of further research aimed at refining these methods. Such refinements would enhance their applicability across a broader range of network environments, thereby improving overall threat detection capabilities.

1.2 Main Contributions

This dissertation presents three primary contributions to the field of cybersecurity.

Firstly, despite extensive research utilizing darknet data, many studies rely heavily on statistical methods that necessitate expert interpretation, limiting their scalability to more comprehensive challenges. These retrospective approaches typically analyze historical darknet data to elucidate past events on specific segments of the Internet. This dissertation addresses these limitations by deconstructing the temporal change detection problem into two phases: clustering similar threat actors and subsequently detecting temporal differences. This automated solution effectively manages the vast amounts of streaming traffic data in near real-time, offering a contemporary perspective on the current threat landscape. This approach facilitates comparison with earlier data to identify new developments, equipping analysts with the capability to anticipate imminent threats and implement proactive measures for threat mitigation.

Secondly, the feasibility of integrating two distinct monitoring systems is explored to leverage the strengths of each while minimizing their respective weaknesses. The dissertation demonstrates how a Network Telescope can enhance threat intelligence gathered by smaller sensors, with the potential for extending this methodology to other sensors and domains with appropriate modifications. Future research may delve deeper into the relationships between these traffic data sources, contributing to a better understanding of malware characteristics across varied systems.

Lastly, the potential for utilizing privileged information during model training to establish meaningful associations between features and labels is investigated. This innovative approach enhances the robustness of data usage in cybersecurity by integrating information from multiple sources.

The findings of this dissertation hold significant promise for advancing cybersecurity systems, including the development of early warning systems and threat intelligence feeds.

By fostering situational awareness, these systems enable cybersecurity analysts to receive timely alerts about imminent threats, thereby empowering informed decision-making and the formulation of proactive strategies for risk anticipation, prevention, and mitigation.

1.3 Research Overview and Organization

In this dissertation, three studies are conducted. The first focuses on developing a temporal change detection mechanism to identify new events within the expansive Internet threat landscape. The second study explores the utility of darknets as a means of inferring threats at an earlier stage by leveraging their capacity as a threat intelligence amplifier. The third study investigates the potential of incorporating information from various sources as privileged data for models designed to analyze darknet data.

The chapters of this dissertation are outlined below, with a summary of the respective studies provided for clarity. Each study contributes to a comprehensive understanding of the dynamics of threat detection and the strategic utilization of darknet data in enhancing cybersecurity measures.

Literature Review

Chapter 2 provides a comprehensive review of the existing literature concerning the utilization of darknets and the synergistic application of darknet data alongside other data sources. The primary objective of this chapter is to identify the challenges and gaps within the current body of research. In this review, the multifaceted roles that darknets play in threat intelligence are examined, highlighting their potential to enhance understanding of malicious activities on the internet. The exploration includes an analysis of how integrating darknet data with additional information sources can yield richer insights into threat landscapes. Additionally, the chapter addresses the inherent limitations and obstacles faced in the effective application of such integrated approaches.

Conceptualization and Definition

This chapter establishes the foundational terminology and technologies pertinent to the research, providing clarity and context for subsequent analyses. A comprehensive description of the data utilized throughout the studies is presented, along with the methods of data representation employed. Key concepts are defined to ensure a shared

understanding of critical terms relevant to cybersecurity and threat intelligence. Furthermore, the chapter elaborates on the types of data collected, including specifics regarding its sources, characteristics, and structure.

Temporal Change Detection in Scanning Activities

In Chapter 4, a novel change detection mechanism is demonstrated, utilizing clustering as a foundational approach. While clustering darknet events effectively groups similar occurrences, it does not provide insights into the temporal evolution of these groups. To address this limitation, optimal transport theory is employed to facilitate a clustering-based temporal change detection framework. This chapter elucidates how the application of optimal transport theory can enhance the analysis of clustering outcomes by providing a mathematical framework for comparing the distributions of clustered events over time. By leveraging this theory, the dynamic nature of event groups can be captured, enabling a more comprehensive understanding of how threats evolve within the darknet landscape.

IP Threat Intelligence Enhancement

In Chapter 5, the potential of leveraging darknets as a threat intelligence amplifier is explored by examining the associations between darknet features and threat intelligence gathered from smaller sensor networks. This chapter investigates how the rich data provided by darknets can enhance the understanding of emerging threats when integrated with insights from other, less extensive surveillance systems. The analysis focuses on identifying key correlations that exist between the characteristics observed in darknet data and the threat intelligence reported by these smaller sensor networks. By establishing these connections, the study aims to demonstrate how darknet data can augment existing threat detection frameworks, thereby providing a more comprehensive view of the threat landscape.

Learning Using Privileged Information

In Chapter 6, the feasibility of utilizing data from various external sources as privileged information is examined to enhance the performance of the associations established in Chapter 5. This chapter delves into how integrating supplementary data can provide additional context and depth to the relationships identified between darknet features and threat intelligence. The study investigates the mechanisms through which this external

data can enrich the existing model, thereby improving its predictive accuracy and overall effectiveness in threat detection. By leveraging diverse datasets, the potential to uncover latent patterns and correlations that may not be apparent when relying solely on darknet information is explored.

Discussion and Future Work

This chapter provides a comprehensive summary of the discussions derived from all three studies presented in the dissertation. Each study's contributions are synthesized to highlight the overarching themes and insights gained throughout the research. This chapter addresses the limitations of this research and motivates future works in this research area.

Conclusion

The final chapter in this dissertation summarizes all three studies.

Chapter 2
Literature Review

2.1 Network Telescopes

Early studies with network telescopes primarily focused on the practical implementations of darknet, but research has since broadened to encompass sophisticated cybersecurity applications, including threat profiling, anomaly detection, and the study of threat variants [20]. This shift reflects a growing understanding of the potential of darknets to provide actionable intelligence. Initial work often leveraged statistical methods and time-series analysis to characterize threats observed within darknet traffic [25, 26]. For example, Harder et al. [20] demonstrated, through a three-month study of a Class C darknet, that a significant portion of observed traffic originated from and terminated at a relatively small number of IP addresses. Furthermore, various techniques have been developed to detect anomalous behaviors and intrusions within these networks [27–29].

Darknet data has emerged as a critical resource in cybersecurity threat analysis, offering invaluable insights into a wide range of network anomalies. These include, but are not limited to, port scans [14–16], worm propagation events [17–21], denial-of-service (DoS) attacks, and various miscellaneous incidents such as misconfigurations [22], internet outages [23, 24], and politically motivated cyber activity [23]. Comprehensive darknet traffic analysis facilitates the rapid identification and mitigation of malicious activities, proving particularly useful in the aftermath of security breaches (e.g., malware distribution, hacking, online fraud) by aiding in attack reconstruction and tracing origins. Furthermore, meticulous analysis provides a nuanced understanding of both legitimate user behavior and the tactics, techniques, and procedures (TTPs) employed by malicious actors (note that terminology for observed IP addresses varies contextually, with terms like “scanners,” “actors” and “source IPs” all being used). This enhanced understanding directly informs the development of effective cybersecurity strategies and policies.

Despite data limitations, darknet data holds immense value for cybersecurity research due to the insights into abnormal network behaviors. Its passive measurement approach has significantly advanced various cybersecurity tasks, including the identification and tracking of novel attack vectors, tracing attack origins, modeling and identifying probing activities, monitoring remote network events, correlating scanning activity to identify coordinated attacks, and assessing the overall cybersecurity hygiene of internet-dependent systems.

Recent research indicates that threat actors frequently conduct reconnaissance campaigns, identifying vulnerable hosts and services prior to launching attacks [30, 31]. Darknets prove exceptionally valuable in detecting such malicious scans and modeling probing behavior, often preceding large-scale coordinated attacks [31]. A significant portion of darknet traffic comprises these probes, which have been extensively studied [32–34]. Furthermore, analysis of spoofed packets on darknets provides vital information for identifying potential victims of spoofing attacks [32–34]. These observations have major implications for enhancing the security of networked systems.

Threat intelligence derived from darknet analysis has substantially enhanced the understanding of organizational cybersecurity posture. Network telescopes have proven crucial in capturing threat landscapes during major events, such as the propagation of prominent worms and botnets like Code Red, Sapphire [19], Witty [35], and Mirai [36]. The scale and heterogeneity of darknet traffic facilitates the detection of large-scale attacks targeting specific systems, such as Internet of Things (IoT) devices [37] and industrial control systems (ICS) [38], enabling timely alerts. Darknet's value is further underscored by its ability to identify victims of denial-of-service (DoS) attacks and to analyze the impact of vulnerability disclosures [39]. Research has demonstrated its utility in tracking threat actors and their infrastructure, highlighting its ongoing significance in cybersecurity.

Despite its potential, harnessing darknet data for threat intelligence faces considerable challenges. The sheer volume and diversity of raw data demand scalable, real-time processing and analysis capabilities to meet the needs of security-sensitive applications. The continuous increase in darknet traffic necessitates even more robust and efficient solutions. The presence of noise, including non-malicious traffic, requires tools capable of distinguishing malicious threats from benign activity. Further complicating the analysis, the inherent nature of data collection can compromise its quality, and the lack of ground truth introduces challenges in evaluating the results of any analysis carried out on top of darknet data. Addressing these issues demands advanced analytic techniques,

machine learning algorithms, and sophisticated data processing tools, coupled with a comprehensive understanding of network infrastructure and cyber threats.

2.2 Darknet Data Representation

The analysis of high-dimensional darknet traffic data presents significant challenges for traditional statistical and machine learning methods. Effective dimensionality reduction is crucial to overcome the "curse of dimensionality" and enable efficient and accurate analysis. Existing literature explores various approaches to represent darknet traffic data, ranging from simple feature extraction to sophisticated deep learning techniques. This review examines several key contributions in this area.

Early work focused on leveraging readily available features from network traffic. Statistical features such as packet header information provide a basic, albeit potentially low-fidelity, representation. However, the complexity and heterogeneity of darknet activity necessitate more advanced representations capable of capturing nuanced patterns.

Dimensionality reduction techniques have proven particularly valuable. For instance, Pour et al. [40] employed L1-norm Principal Component Analysis (PCA) to reduce the dimensionality of network telescope data, enabling the inference of coordinated Internet of Things (IoT) scanning campaigns. This approach effectively highlights the principal components of the observed activities, facilitating the identification of coordinated malicious behavior. Similarly, Cabana et al. [38] combined PCA with clustering and graph-based analytics to analyze scanning data targeting industrial control systems. Their work demonstrates the synergy between dimensionality reduction and other analytical techniques in uncovering the source and nature of targeted attacks. This approach focuses on extracting meaningful features from payload inspection to enhance classification accuracy.

More recently, deep learning methods have emerged as powerful tools for learning complex, non-linear representations of darknet traffic. Autoencoders, in particular, have demonstrated significant promise. Sarabi et al. [41] showcased the versatility of well-trained autoencoders in generating low-dimensional representations of active internet-wide scanning data from Censys. These compact representations proved effective for various tasks, including the detection and prediction of malicious hosts, demonstrating the generalizability of the learned features. Furthermore, Kallitsis et al. [42] successfully employed autoencoders for representing traffic data from the ORION darknet. Their work highlights the ability of autoencoders to capture inherent patterns within the

high-dimensional data, facilitating subsequent clustering and temporal event detection. This approach, which focuses on learning an information-preserving low-dimensional embedding, forms the foundation of this research study. The present research builds upon this foundation, employing autoencoders to generate low-dimensional representations of high-dimensional scanning data for subsequent analysis.

2.3 Integration with Other sources

The escalating sophistication of cyber threats necessitates a robust, integrated approach to threat intelligence gathering that transcends the limitations of individual data sources. While darknet monitoring and honeypot deployments each provide valuable insights, their isolated analysis yields an incomplete understanding of threat actors and their tactics, techniques, and procedures (TTPs). This section of literature review examines current methodologies integrating darknet and honeypot data, identifies their limitations, and proposes avenues for advancement using supervised machine learning.

Early research investigated probing activities using individual sensors—network telescopes, honeypots, and intrusion detection systems (IDS)—within specific network contexts [43–45]. These studies, often conducted within large campus networks [43–45], characterized scanning behaviors and assessed associated risks. Concurrent analyses of darknet traffic provided a broader perspective on internet-wide scanning events, enriching threat intelligence related to targeted services. However, the inherent limitations of relying on single data sources, primarily the substantial data volume demanding extensive manual analysis, hinder the efficient generation of actionable intelligence. This necessitates the development of integrated analytical approaches.

Integrating darknet and honeypot data offers a potential solution. Akiyoshi et al. [46] demonstrated the efficacy of a hybrid system combining low-interaction honeypots and darknet monitoring for automated reconnaissance campaign detection, leveraging the complementary strengths of both: darknet's macroscopic view of large-scale incidents and honeypots' microscopic view of attacker behaviors. Subsequent research correlated network telescope and honeypot data to quantify attacks targeting specific protocols [47, 48], yet these studies predominantly focused on specific attack vectors and often lacked precise attribution capabilities.

Despite advancements in integrating network telescope and honeypot data for improved threat detection and quantification, the potential for enhanced IPv4 threat intelligence remains largely unrealized. This gap motivates the exploration of more so-

phisticated analytical techniques, specifically supervised machine learning. By leveraging the richness of multi-source data, including threat labels from honeypot data, supervised learning models can identify complex patterns within high-dimensional network traffic data, thereby improving threat detection and attribution [49].

Supervised learning has demonstrated considerable efficacy in various network traffic analysis tasks, including traffic classification [50], anomaly detection [51], and network performance monitoring [52]. However, applying these techniques to large-scale internet traffic data presents challenges. The high dimensionality of this data, with scanners probing tens of thousands of ports daily [53], necessitates careful feature engineering and selection to mitigate the curse of dimensionality and ensure effective model training. Future research should therefore prioritize the development of robust feature extraction methods capable of capturing intricate relationships within high-dimensional data, thereby improving the accuracy and scalability of supervised learning-based threat intelligence systems.

Chapter 3
Conceptualization and Definition

3.1 Background

Network telescopes or darknets, are passive observation systems strategically deployed across the internet infrastructure to capture and analyze anomalous network traffic [54]. These systems, sometimes referred to as network sinks, blackhole monitors, or packet telescopes, monitor a designated portion of the IP address space, often termed the "dark IP space," which is not assigned for legitimate network services [12,13]. Traffic directed to this unassigned space, commonly known as Internet Background Radiation (IBR) [12,13], is highly suspicious and warrants detailed investigation. This IBR comprises unsolicited probes from botnets and network scanning tools, malware propagation attempts, denial-of-service (DoS) attack backscatter, and traffic originating from misconfigured network devices. The inherent isolation of malicious activities within IBR provides a significantly enhanced signal-to-noise ratio compared to the analysis of general internet traffic.

The non-intrusive nature of network telescopes is a key advantage. Operating passively, these systems collect data without interfering with normal internet functionality. This unidirectional data flow, however, presents a limitation: the absence of significant payload data in the predominantly TCP SYN packets received limits forensic analysis capabilities. While the lack of a completed TCP handshake prevents the interception of post-handshake payload data, the analysis of observed connection attempts and associated metadata remains valuable for characterizing network threats and attacker behavior.

Traditional network intrusion detection systems (NIDS) often rely on threshold-based approaches to detect scanning activity. These approaches typically establish thresholds on the number of packets from a suspect host within a defined timeframe or the number of unique destinations contacted (e.g., 25 unique destinations within 5 minutes). While effective in identifying high-intensity scanning activities, these methods struggle to detect

low-frequency, stealthy attacks. Lowering the thresholds to increase sensitivity inevitably leads to an increased false positive rate, overwhelming analysts with alerts and escalating the complexity of distinguishing malicious from benign events. The inherent isolation of malicious activities within the IBR observed by network telescopes addresses this limitation. Because benign user traffic is largely absent from the darknet, the need for arbitrary thresholding is eliminated. This allows for the detection of even low-intensity scanning activities, providing a significant advantage over traditional methods [54].

The effectiveness of data collection in network telescopes is closely tied to their scale and deployment strategy. While small-scale telescopes are usually confined to individual locations for experimental purposes, medium-scale setups can span entire regions or countries, enabling targeted threat monitoring. In stark contrast, large-scale, globally distributed telescopes excel in providing detailed resolution and broad coverage. This results in extensive datasets that allow for intricate analyses of internet traffic patterns. The ability of larger darknets to detect low-level, low-frequency traffic is particularly noteworthy, as such traffic often signals covert and prolonged attacks—something that smaller systems struggle to identify. Moreover, the long-term monitoring capabilities of these large-scale telescopes are crucial for recognizing these slow-moving threats. Importantly, the passive, non-invasive nature of network telescopes ensures that regular internet operations remain undisturbed, thus alleviating potential legal challenges associated with their use.

3.1.1 ORION Network Telescope

The ORION network telescope (shown in Figure 3.1), operated by Merit Network, Inc., passively monitors a substantial segment of the IP address space. It encompasses 1,856 /24 subnets, which collectively represent an estimated 500,000 unique IP addresses within this hidden network. This sophisticated infrastructure facilitates the ongoing collection and logging of network traffic, yielding a voluminous dataset that surpasses 2 billion packets daily, originating from more than 650,000 distinct source IP addresses. Furthermore, the telescope's data processing pipeline is designed to detect indicators of compromise (IOCs) that are commonly associated with malicious network scanning activities. This capability enhances the system's effectiveness in identifying potential threats within the extensive data it gathers.

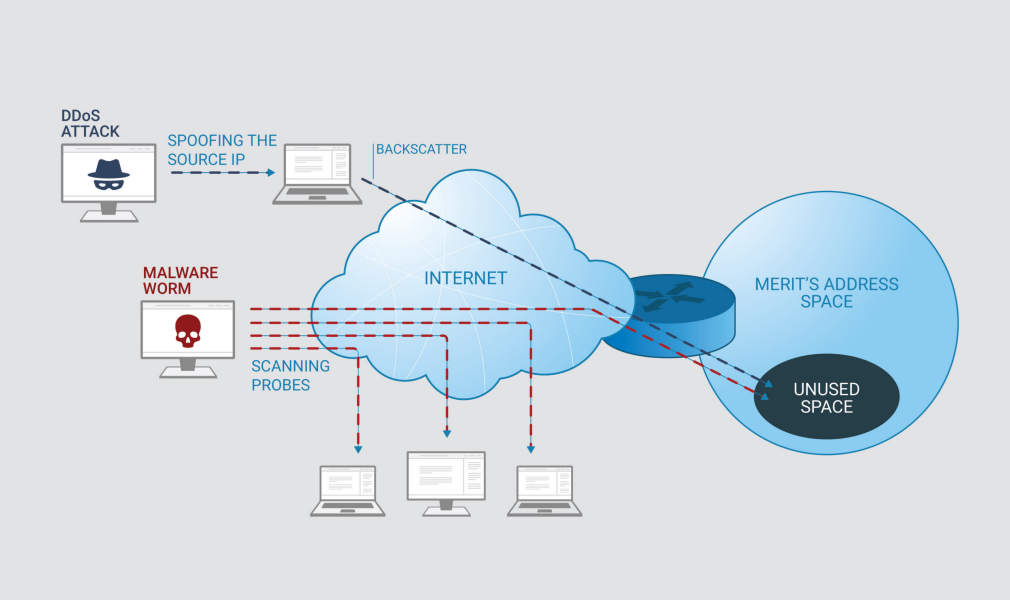


Figure 3.1: Scanning and backscatter traffic captured in the Darknet [1].

3.2 Darknet Data

The ORION network telescope utilizes a near real-time data acquisition pipeline to effectively extract and annotate scanner activities. Every hour, packets from the designated dark IP space are captured in PCAP (Packet Capture) format, resulting in a substantial daily collection of over 100 GB of darknet data. This raw PCAP data undergoes a comprehensive processing phase aimed at identifying significant darknet events, such as scanning and backscatter activities.

To enhance the analysis, these events are supplemented with external data sources, including DNS information and geolocation data from MaxMind databases [55], as well as routing details derived from CAIDA's prefix-to-Autonomous System mappings dataset (pfx2as) [56]. Each darknet event is characterized by specific attributes, including its source IP address, protocol flags, and targeted ports. A caching mechanism is employed to keep track of active scanners and their associated events in memory, while inactive events—defined as those with no recorded activity for approximately 10 minutes—are archived to disk. Importantly, scanners that target multiple ports and/or protocols are treated as distinct events, a consideration that plays a vital role in the feature engineering process.

To facilitate efficient analysis and integration with supplementary datasets, all identified darknet events are uploaded to Google BigQuery in near real-time. This integration

allows for the incorporation of data from Censys [57], which conducts comprehensive scans of the entire IPv4 address space. The information obtained from Censys is invaluable, as it provides contextual details about scanning hosts, including open ports and active services. The benefit of this integration is demonstrated in [36], where Censys data was instrumental in identifying device types and manufacturer information associated with the Mirai botnet. In this research, Censys data is used to enhance the feature set used in clustering analyses. Daily aggregation of darknet data alongside integrated external data sources forms the baseline for clustering darknet events.

The ORION pipeline emphasizes the identification of indicators of compromise (IOCs) by employing metrics that quantify the frequency, scope, and intensity of scanning behaviors. This analysis is further refined through the inclusion of packet header metadata, such as Time-To-Live (TTL) and IP Identification (IPID) values, thereby bolstering IOC detection capabilities. These enhancements support the development and evaluation of advanced anomaly and intrusion detection systems specifically designed to address the unique characteristics inherent in the darknet environment.

3.2.1 Data Preprocessing and Aggregation

Raw darknet events data is aggregated daily to streamline the analysis process. Each scanner detected within a given day, irrespective of the number of ports or protocols it scans, is represented as a singular data point. This aggregation method allows for an effective examination of daily scanner behavior while minimizing the influence of high-frequency, low-level fluctuations in scanning activity. Each aggregated data point is characterized by a comprehensive feature set that includes both numerical and categorical attributes, enhancing the robustness of the analysis.

A scanning profile for each sender is constructed by aggregating relevant information pertaining to that sender over a 24-hour period. This consolidated dataset serves as the core features for the scanner profile. The subsequent subsections outline the features chosen for this research and provide the rationale for their selection.

3.2.2 Category of Features

3.2.2.1 Network Activity Metrics

Key indicators of a scanning actor's intensity and strategy include the *number of packets* transmitted, the *total volume of bytes* sent, and the *inter-arrival time* of packets. An

elevated count of packets and bytes, combined with brief inter-arrival times, frequently correlates with aggressive scanning behaviors that may suggest malicious intent. These features indicate a rapid and sustained scanning activity directed towards identifying vulnerable targets.

3.2.2.2 Scanning Strategy

The *number of distinct ports* and *destination IP addresses* accessed by a scanner serves as a critical indicator of the scanning strategy utilized to identify potential targets. Large darknets provide a valuable perspective for observing contiguous scanning activities and deciphering these strategies, thereby advancing our understanding of the ultimate objectives behind scanning campaigns.

In addition, several key fields are defined, including *prefix density*, *destination strategy*, *IPID strategy*, and *IPID options*, which collectively encapsulate the probing methodology. *Prefix density* is quantified as the ratio of the number of scanners within a given routing prefix to the total number of IP addresses encompassed by that prefix. This metric, which utilizes CAIDA's pfx2as dataset for mapping IPs to their corresponding routing prefixes, offers insights into the orchestration of scanning efforts within a network.

The *destination strategy* and *IPID strategy* are features that reflect the states of the associated fields, specifically destination IP and IPID, respectively. These strategies may involve the scanning entity either 1) maintaining a constant value, 2) incrementing in fixed steps, or 3) randomizing across consecutive probes. Such features can yield insights into the scanning tools employed; for example, the ZMap tool is known to utilize a constant IPID value of 54321, which can illuminate the intentions behind the scanning activity.

Finally, the *TCP options* field is represented as a binary feature indicating whether any TCP options have been configured during TCP-related scanning. The absence of TCP options has been linked to aggressive and hostile scanning strategies.

3.2.2.3 Scanning Device

The *Time-To-Live (TTL) values* serve as important indicators for identifying the operating system (OS) type of devices [58] and can also be indicative of "irregular scan traffic" [59,60]. Specifically, packets originating from Unix-based operating systems typically exhibit TTL values ranging from 40 to 60, as these systems are initialized with a default TTL of 64. Conversely, devices running Windows OS are assigned an initial TTL value of 128, which translates to captured values between 100 and 128 in darknet observations. Given

that many attacks are tailored to specific devices, the accurate identification of targeted system types is essential for mitigating potential threats effectively.

3.2.3 Assumptions

To facilitate the approximation of scanning behaviors and to streamline the creation of scanner profiles for subsequent analysis, certain foundational assumptions are established. A pivotal assumption employed throughout this research posits that if an IP address is detected within any location during the designated 24-hour observation window, it is inferred that the actor associated with this IP address remains constant. This assumption is generally valid, as the sensors in question are positioned in areas where the occurrence of legitimate traffic is markedly low. By concentrating on these less trafficked regions of the network, the likelihood of encountering benign interactions is significantly diminished, thereby enhancing the reliability of the data collected for analysis. Such assumptions are essential for drawing meaningful conclusions from the scanning activities observed and for developing robust profiles that accurately reflect the underlying malicious intent.

3.3 Data Representation

The efficacy of machine learning (ML) algorithms is intrinsically linked to the chosen data representation, as the representation encapsulates the underlying factors driving data variability [61]. Dimensionality reduction techniques are crucial for mitigating the computational challenges posed by high-dimensional data while preserving essential information. Principal Component Analysis (PCA), a widely adopted linear dimensionality reduction method, projects data onto a new coordinate system where a reduced number of principal components capture the majority of data variance [62]. While Fukuda et al. [63] demonstrated the sufficiency of the first four principal components to characterize traffic behavior variations, PCA's linearity limits its capacity to model the often prevalent non-linear relationships between traffic features. In response to these limitations, alternative methodologies have emerged, including the Fourier Transform and Kalman Filtering techniques as applied in [31], which aim to derive latent space representations from time-series data without the necessity for direct processing of raw traffic data.

Recent advancements in domain-specific techniques have further contributed to the encoding of high-dimensional traffic data vectors. For instance, IP2Vec [64] adapts the

principles of Word2Vec [65] to create vector representations for IP addresses, embedding contextual behavioral data extracted from flow information. This method ensures that IP addresses with similar behavioral patterns yield vectors that exhibit high cosine similarity. Similarly, DANTE [66] employs a comparable vector representation framework to categorize Internet hosts with analogous behaviors. In contrast, DarkVec [67] presents a modified Word2Vec approach that focuses on co-occurring source-destination port access patterns, thereby capturing similar port-scanning behaviors within the darknet.

However, it is important to note that the specialized and task-oriented nature of these domain-specific representations limits their applicability across different domains or tasks. Additionally, there exists a substantial risk of overfitting to the training data, which may compromise the model's ability to generalize effectively to previously unseen data.

Building upon these insights and following extensive experimentation with various representation methods, the decision was made to utilize autoencoders for learning latent representations of darknet data. Autoencoders employ non-linear activation functions in their hidden layers, allowing them to capture complex, non-linear relationships between the input data and the encoded outputs. The resulting compressed representation serves as a universal, low-dimensional vector that effectively encapsulates the high-dimensional and heterogeneous features of network traffic. This versatile representation can be subsequently leveraged for a range of downstream tasks, including classification, clustering, and anomaly detection.

Deep autoencoders [68–70] have garnered significant attention within the field of deep learning for their ability to produce latent representations of data [61, 69]. A non-linear autoencoder is fundamentally composed of two essential components: the encoder function, denoted as $\theta(\cdot)$, which is parameterized by θ and is responsible for mapping the input space $\mathcal{X} \in \mathbb{R}^P$ to a latent representation space $\mathcal{Z} \in \mathbb{R}^Q$, and the decoder function, represented as $\mu(\cdot)$, parameterized by μ, which maps the latent space $\mathcal{Z}$ back to the original data space $\mathcal{X}$.

Through this process, the autoencoder effectively compresses a high-dimensional input signal $\mathbf{x} \in \mathbb{R}^P$ into a lower-dimensional embedding $\mathbf{z} \in \mathbb{R}^Q$, where it is ensured that $Q \ll P$. The overarching objective is to maintain as much information as possible throughout this transformation (see Figure 3.2). Mathematically, these mappings can be articulated as follows:

$$\mathbf{z}_i = \theta(\mathbf{x}_i) = f(\mathbf{x}_i; \theta) \quad f(\cdot; \theta) : \mathbb{R}^P \to \mathbb{R}^Q \tag{3.1}$$

$$\hat{\mathbf{x}}_i = \mu(\mathbf{z}_i) = g(\mathbf{z}_i; \mu) \quad g(\cdot; \mu) : \mathbb{R}^Q \to \mathbb{R}^P \tag{3.2}$$

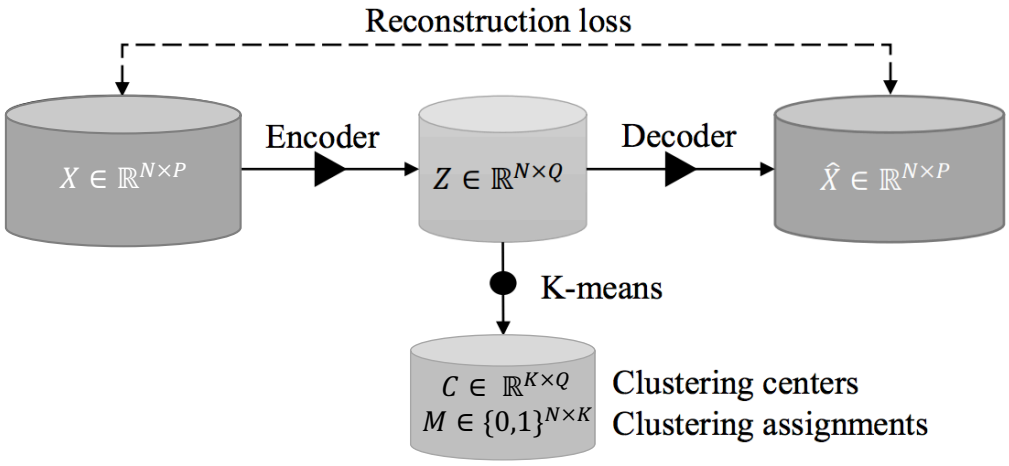


Figure 3.2: Autoencoder Architecture for Dimensionality Reduction

The learning process of the autoencoder is driven by the minimization of the reconstruction error, which measures the disparity between the original input and the corresponding decoded output. This objective function serves to quantify the differences between the input $\mathbf{x}_i$ and the reconstructed data $\hat{\mathbf{x}}_i$, thereby ensuring that the representation learned effectively encapsulates the essential features of the data. The optimization problem can be formally expressed as follows:

$$\min_{\theta,\mu} \sum_{i=1}^{N} \ell(g(f(\mathbf{x}_i; \theta); \mu), \mathbf{x}_i) + \lambda(R(\theta) + R(\mu)) \tag{3.3}$$

In this context, the reconstruction error is defined as $\ell(\cdot) : \mathbb{R}^P \to \mathbb{R}$, calculated using the squared ℓ_2 norm, expressed mathematically as $\ell(\mathbf{x}, \mathbf{m}) = |\mathbf{x} - \mathbf{m}|_2^2$. Regularization terms, denoted as $R(\cdot)$, are incorporated to penalize the complexity of the model, thereby mitigating the risk of overfitting. Specifically, ℓ_2 regularization is employed, represented as $R(\theta) = |\theta|_2^2$ and $R(\mu) = |\mu|_2^2$, with a regularization strength controlled by the parameter $\lambda \geq 0$.

The functions $f(\cdot; \theta)$ and $g(\cdot; \mu)$ are realized through fully-connected multilayer perceptrons (MLPs). As an example, a 4-layer MLP encoder $f(\cdot; \theta)$ can be defined as follows:

$$
\begin{aligned}
\mathbf{h}^{(1)} &= \phi(\mathbf{W}^{(1)}\mathbf{x} + \mathbf{b}^{(1)}) \\
\mathbf{h}^{(2)} &= \phi(\mathbf{W}^{(2)}\mathbf{h}^{(1)} + \mathbf{b}^{(2)}) \\
\mathbf{y} &= \phi(\mathbf{W}^{(3)}\mathbf{h}^{(2)} + \mathbf{b}^{(3)}) = f(\mathbf{x}; \theta)
\end{aligned}
$$

In this formulation, the element-wise ReLU activation function, denoted as $\phi(\cdot)$, is employed to address the vanishing gradient issue that can arise with sigmoid functions. The parameters $\mathbf{W}$ and $\mathbf{b}$ represent the learned weight matrices and bias vectors, respectively, which are optimized through backpropagation and stochastic gradient descent. The notation $\mathbf{h}^{(\ell)}$ signifies the output of the ℓ-th hidden layer within the architecture.

The decoder function $g(\cdot; \mu)$ is designed to mirror the encoder's structure, maintaining symmetry in both the number and size of the hidden layers. Similar to the encoder, the ReLU activation function is applied across all hidden layers, with the exception of the final layer, which employs a linear activation to ensure that the output remains within the same space as the input, $\mathbb{R}^P$.

Chapter 4
Temporal Change Detection in Scanning Activities

This chapter investigates the complexities inherent in threat detection utilizing darknet data, presenting a robust mechanism capable of effectively tracking temporal fluctuations within the Internet threat landscape. Through this framework, the research successfully elucidates methods for identifying emerging malicious events, thereby enhancing our understanding of evolving cyber threats. By analyzing the dynamic interactions within the darknet, the study underscores the significance of continuous monitoring and adaptive response strategies in the ongoing battle against cybercrime. The findings emphasize the necessity for advanced analytical tools that can not only detect but also anticipate potential threats.

4.1 Background

The analysis of network traffic for malicious activity presents significant challenges. The inherent difficulty in discerning malicious traffic from the overwhelming volume of legitimate network activity is exacerbated by sophisticated obfuscation and encryption techniques employed by malicious actors. Furthermore, privacy concerns often restrict data sharing, limiting the availability of sufficiently large datasets of malicious traffic for comprehensive analysis. This data scarcity problem is particularly acute when studying specific types of attacks or malware.

In contrast, large-scale network telescopes offer a unique advantage. Their wide aperture and persistent surveillance capture a higher concentration of malicious traffic due to the relatively low volume of legitimate traffic observed. This allows for high-resolution temporal analysis of low-frequency traffic patterns often associated with covert

malicious activities, such as Advanced Persistent Threats (APTs) and botnets, which are frequently missed by other monitoring systems. The temporal dimension is crucial; analyzing changes in traffic patterns provides an opportunity to detect emerging threats, including those designed to evade conventional detection methods. This approach enables identification of novel malware variants by tracking fluctuations in associated traffic behavior and assessment of the impact of security events, such as vulnerability disclosures, through the analysis of subsequent traffic pattern changes. A sudden increase in traffic associated with a specific threat following a vulnerability announcement provides strong evidence of exploitation. Moreover, the identification of sudden alterations in traffic patterns linked to specific source IP addresses, geographic locations, protocols, or services enables threat attribution. The broader coverage of network telescopes increases the likelihood of early threat detection compared to other monitoring systems, providing a crucial early warning capability for cybersecurity professionals.

4.2 Temporal Change Detection

Temporal change detection (TCD) constitutes a critical methodology for identifying and analyzing alterations in data patterns over time. Within the cybersecurity domain, TCD plays a crucial role in monitoring network traffic dynamics to detect anomalies indicative of potential security breaches. This involves the continuous acquisition and analysis of network traffic data over a defined period, comparing observed patterns against an established baseline to identify abrupt deviations in traffic volume, type, or other relevant metrics. Such deviations can signal a range of threats, from denial-of-service (DoS) attacks to the infiltration of malicious code. Furthermore, TCD can effectively pinpoint performance bottlenecks or congestion within a network, potentially revealing vulnerabilities exploitable by adversaries. Addressing these vulnerabilities through proactive TCD enhances overall network security and resilience.

The importance of TCD extends beyond cybersecurity. Its utility has driven extensive research across diverse fields, including finance (detecting market shifts), healthcare (identifying changes in patient health indicators), social media analysis (monitoring evolving user behavior), and environmental monitoring (tracking environmental condition changes). The proliferation of time series data in these domains has stimulated the development of sophisticated TCD algorithms, ranging from basic statistical methods to advanced machine learning (ML) models. These methods are designed to discern subtle shifts in data patterns amidst noise and stochastic fluctuations. The continued

exponential growth of time-series data underscores the growing importance of robust and efficient TCD techniques across a multitude of application areas.

4.2.1 Taxonomy

TCD algorithms can be classified along several axes. One primary categorization distinguishes between *offline* and *online* techniques. Offline methods assume the availability of the entire dataset prior to analysis, enabling comprehensive examination to identify temporal changes. Statistical approaches such as regression analysis, time series decomposition, and principal component analysis exemplify this category. Conversely, online techniques process data streams in real time, detecting changes as they occur. This is crucial for applications demanding immediate responsiveness, such as real-time network intrusion detection. Examples include cumulative sum (CUSUM), exponentially weighted moving average (EWMA), and sequential hypothesis testing (SHT) [71].

Another classification distinguishes *parametric* and *non-parametric* methods. Parametric methods assume the data follows a specific probability distribution, estimating model parameters to detect shifts in distributional properties like mean and variance. Examples include autoregressive integrated moving average (ARIMA) models, exponential smoothing (ES), and generalized linear models (GLMs). Non-parametric methods, in contrast, make no distributional assumptions, often employing rank-based statistics such as the Wilcoxon-Mann-Whitney test, the Kolmogorov-Smirnov test, or the CUSUM algorithm to detect changes. These are particularly useful when dealing with data exhibiting non-normality or outliers.

Finally, TCD methods can be broadly categorized as *statistical* or *machine learning-based*. Statistical techniques leverage probability distributions and statistical tests, whereas ML methods learn from data without explicit programming. ML approaches further subdivide into supervised and unsupervised learning, depending on the availability of labeled training data.

4.2.2 Change Point Detection in Cybersecurity

A common application of TCD in cybersecurity involves change point detection (CPD), focused on identifying specific time instances where abrupt changes occur in the data. Such abrupt shifts often indicate significant alterations in the data-generating process, frequently signaling the initiation of malicious activities. While TCD encompasses the broader detection of changes in time series properties, CPD specifically targets pinpointing

the exact times of significant changes. Numerous statistical and ML algorithms have been adapted for CPD, primarily comparing probability distributions across time intervals to identify significant shifts.

Traditional CPD methods often rely on the likelihood ratio, comparing the probability densities of consecutive intervals. Subspace modeling provides an alternative approach, representing time series using state spaces and using this representation to predict state space parameters for detecting change points. Probabilistic methods estimate probability distributions of new intervals based on previously observed data. Kernel-based methods map observations to higher-dimensional feature spaces to assess subsequence homogeneity. Graph-based techniques represent time series as graphs, applying statistical tests to detect changes in the graph structure. Finally, clustering-based methods group time series data, identifying changes by comparing the features of the resulting clusters.

The CUSUM algorithm, originally introduced in [72], enjoys widespread use in network anomaly detection. This algorithm computes the cumulative sum of deviations from a reference value, flagging a change when the cumulative sum exceeds a predetermined threshold. Wang et al. [73] and Siris et al. [74] employed CUSUM to detect SYN flooding DoS attacks by identifying change points in the cumulative number of SYN packets received. Non-parametric CUSUM has also been applied to worm detection [75,76], based on the number of probed destination hosts. Ahmed et al. [77] developed an adaptive, sliding-window CUSUM algorithm for automatically detecting changes in network traffic parameters, further enhanced in [78] with dynamic sliding windows to handle nested changes.

Inoue et al. [79] proposed a CPD method that learns a statistical model from time-series data, calculating anomaly scores for each time instance and combining them into a single metric reflecting the likelihood of a change point. This approach demonstrated success in identifying worm outbreaks and large-scale DDoS attacks. Sun et al. [80] presented a Bayesian inference-based method for change point detection, serving as a complementary tool to improve the robustness of their probe detection system.

4.2.3 Challenges in Temporal Change Detection of Darknet Traffic

While the rich dataset of malicious traffic available within darknet presents a valuable resource for threat analysis, several challenges impede near real-time threat identification for practical cybersecurity applications. The sheer volume of data necessitates efficient processing techniques, while the intricate and multifaceted nature of network traffic complicates accurate pattern identification and change detection. The heterogeneity

of traffic types and sources demands robust and adaptable detection methods capable of handling diverse traffic characteristics. The continuous evolution of adversarial tactics necessitates the development of innovative and adaptable detection techniques to counter emerging threats. A further major challenge is the inherent high false positive rate of many change detection algorithms. Careful parameter tuning and threshold selection are therefore crucial for achieving high accuracy and low false positive rates. Finally, understanding the cause of detected changes, rather than simply identifying their occurrence, remains a significant challenge. Many existing change detection algorithms lack the necessary interpretability to provide this crucial contextual information. This necessitates the development of explainable AI techniques that can shed light on the root causes of observed temporal changes in darknet traffic.

4.2.4 Clustering-Based Temporal Change Detection

Many existing CPD methodologies struggle with scalability and interpretability when confronted with massive datasets. Clustering-based methods offer a compelling alternative, focusing on clustered outcomes rather than raw data. These approaches employ a model-fitting paradigm, detecting changes when new data points fail to integrate into existing clusters. This typically involves two overlapping windows: a reference window for cluster creation and a current window for new data. Points not assigned to any cluster are identified as change points. This eliminates the need for explicit thresholds, providing a robust and efficient method for change detection.

Clustering simplifies analysis, facilitating efficient processing of large datasets. Clustering techniques can reveal patterns and changes obscured in raw data, enhancing detection accuracy. Moreover, some clustering-based methods can handle heterogeneous data sources. The combination of clustering and ML algorithms allows for adaptation and learning, improving accuracy over time. Crucially, clustering provides interpretable results, grouping data points into meaningful clusters that allow for insightful understanding of the nature of the detected changes. This interpretability, often lacking in binary-output methods, is especially valuable in applications where understanding the change's nature is paramount. Visualization tools further enhance the interpretability of clustering-based TCD outputs.

This dissertation investigates the application of Optimal Mass Transport (OMT) [81] for real-time change detection in cybersecurity, leveraging clustering techniques to enhance efficiency and scalability. Traditional change detection methods, such as hypothesis testing and time series analysis, often struggle with high-dimensional data and complex

distributions characteristic of modern cybersecurity threats. In contrast, OMT offers a powerful framework for comparing probability distributions, even in high-dimensional spaces, by quantifying the "work" required to transform one distribution into another. This is commonly measured using the Earth Mover's Distance (EMD) [81], a metric sensitive to subtle shifts in data structure.

The proposed methodology encompasses two fundamental stages designed to enhance the efficacy of threat detection. Initially, a computationally efficient clustering algorithm is employed on a refined, low-dimensional representation of the cybersecurity data stream. This dimensionality reduction process is imperative for optimizing computational efficiency and alleviating the challenges posed by the curse of dimensionality, which frequently afflicts high-dimensional datasets typical in cybersecurity contexts. This preprocessing phase converts the raw data into a collection of clusters, each potentially indicative of a unique activity or threat actor. The selection of the clustering algorithm is a pivotal aspect of this methodology; its influence on overall performance is meticulously examined in subsequent sections. By systematically analyzing the clustering outcomes, this research aims to delineate the most effective strategies for identifying and characterizing malicious behavior within the vast expanse of darknet data.

The second stage employs OMT to monitor temporal changes in the cluster distributions. Specifically, we compare the distributions of cluster memberships between consecutive time windows. A significant divergence in these distributions, as quantified by the EMD, signals a change in the underlying activity patterns. This approach avoids making restrictive assumptions about the underlying data distribution, offering robustness against noisy or non-stationary data common in real-world scenarios. Furthermore, the magnitude of the EMD provides a quantitative measure of the change, enabling a more nuanced understanding of the severity and potential impact of detected anomalies.

The advantages of this OMT-based approach are threefold. Firstly, its flexibility and robustness stem from its non-parametric nature, requiring no assumptions about the statistical properties of the data. Secondly, the use of clustering reduces computational complexity compared to directly applying OMT to the raw high-dimensional data, making real-time detection feasible. Thirdly, the cluster-based approach allows for detailed investigation of the characteristics of emerging threats. Newly formed or significantly altered clusters, identified as drivers of the detected changes, can be individually analyzed to understand the underlying threat actor behavior and potentially inform targeted mitigation strategies. This granular analysis provides valuable actionable intelligence beyond simple anomaly detection.

4.3 Clustering

The analysis of darknet traffic presents a significant challenge in modern cybersecurity. The inherent anonymity and dynamic nature of this traffic necessitate advanced analytical techniques to detect malicious activities and identify emerging threats. Unsupervised clustering, a powerful machine learning approach, offers a promising solution by grouping similar traffic patterns based on extracted features, thus enabling the identification of anomalous behavior and facilitating proactive security responses. This section explores the application of various unsupervised clustering algorithms to darknet traffic analysis, highlighting their strengths and limitations in the context of evolving attack methodologies.

Early approaches to darknet traffic clustering leveraged self-organizing maps (SOMs) as a substitute for computationally expensive supervised methods [82]. Subsequently, k-Means and Expectation-Maximization (EM) algorithms were adopted, demonstrating efficacy in certain scenarios [83]. However, these methods often rely on simplistic assumptions about attacker behavior. For instance, Pang et al. [12] proposed grouping events based on average packet delay, assuming consistent fingerprints for individual attackers or malware. This assumption is increasingly invalidated by the sophisticated techniques employed by modern adversaries to introduce stochasticity into their network behavior and evade detection through conventional signature-based methods.

To address the limitations of feature-based clustering, research has shifted towards graph-based representations of darknet activities, offering a more nuanced approach to capturing complex relationships. This shift reflects a recognition that the interdependencies between various network events are often more indicative of malicious activity than individual characteristics alone. Soro et al. [84], for example, constructed a port scan graph embedding port scan frequency and utilized community detection algorithms to identify groups potentially associated with coordinated attacks. This approach leverages the inherent network structure to uncover relationships between seemingly disparate activities and pinpoint common targets of probing activity, providing a more robust and adaptable method for detecting coordinated malicious campaigns. However, the computational complexity of graph-based methods can be considerable, especially for large-scale darknet datasets, requiring optimization techniques and careful consideration of scalability. Further research is needed to address the challenges in balancing the granularity of information captured by graph-based approaches with the computational demands of analyzing massive datasets.

Recent research directions focus on developing hybrid approaches that combine the strengths of feature-based and graph-based methods. This might involve utilizing feature-based clustering to pre-process the data, reducing its dimensionality and computational complexity before employing graph-based methods for higher-level analysis of relationships between identified clusters. Additionally, the exploration of novel unsupervised clustering algorithms designed specifically for the dynamic and heterogeneous nature of darknet traffic warrants further investigation. The development of more robust and scalable techniques is critical to harnessing the full potential of unsupervised learning in mitigating the ever-evolving threats posed by darknet activities.

In this study, the latent space representation of traffic characteristics generated by the "trained" encoding function $f(\cdot;\theta)$ for each observed Internet host is subjected to standard k-means clustering. The objective of this clustering process is to group the data points in $\mathcal{X}$ into a set of k clusters, where k is a user-defined parameter, such that each data point belongs to only one cluster. This optimization problem seeks to minimize a clustering criterion or distance measure, and can be expressed as follows:

4.4 Optimal Mass Transport

Optimal transport (OT) theory [81] provides a rigorous mathematical framework for quantifying the cost of transforming one probability distribution into another. This framework centers on identifying an optimal transportation plan that minimizes the overall cost, typically defined by a distance metric between points in the distributions, subject to constraints ensuring mass conservation. The versatility of OT has led to its widespread adoption in diverse fields, including computer vision (e.g., image registration [81]) and data analysis (e.g., change detection). Within the context of this dissertation, we leverage OT's capabilities for robust cluster comparison.

A prominent distance measure within the OT framework is the Earth Mover's Distance (EMD), also known as the Wasserstein distance. The EMD quantifies the minimum cost to transform one probability distribution into another, providing a measure of dissimilarity between distributions [81]. This characteristic makes the EMD particularly well-suited for tasks requiring the comparison of probability distributions, such as image matching and comparison, where it effectively quantifies the distance between images based on their underlying distributions. In our analysis of evolving Darknet structures, we utilize the EMD to quantify the dissimilarity between clusters across different time points, allowing for the tracking of structural changes.

Our approach uses the Kantorovich formulation of optimal transport (OT), avoiding the more restrictive Monge formulation. The Kantorovich formulation aims to find an optimal transport plan—a probability measure over the product space of two distributions—that minimizes transport costs while ensuring mass conservation. This flexibility allows for non-unique optimal solutions, making it more applicable to complex distributions than the Monge formulation, which requires a one-to-one mapping and often proves impractical.

To compare clusters using OT, we represent each cluster as a probability distribution, employing methods like kernel density estimation. We then calculate the Earth Mover's Distance (EMD) using a chosen metric, such as Euclidean distance, to quantify the dissimilarity between these distributions. This methodology facilitates the systematic tracking of changes in Darknet structure over time, enabling comparative analyses of different Darknet senders and providing a robust framework for assessing the evolution of Darknet topology and actor behavior.

4.4.1 Kantorovich Formulation

In the Kantorovich formulation of the optimal transport problem, two probability density functions (PDFs), I_0 and I_1, are defined over spaces Ω_0 and Ω_1, respectively, where Ω_0 and Ω_1 are typically subspaces in $\mathbb{R}^d$. The aim is to find a transport plan, γ, that transforms I_0 into I_1. This transport plan is a joint probability distribution of I_0 and I_1, and the value of $\gamma(A \times B)$ denotes the amount of mass in set $A \in \Omega_0$ that is transported to set $B \in \Omega_1$. The transport plan γ must satisfy two constraints: (i) $\gamma(\Omega_0 \times B) = I_1(B)$ and $\gamma(A \times \Omega_1) = I_0(A)$, where $I_0(A) = \int_A I_0(x)dx$ and $I_1(B) = \int_B I_1(x)dx$, and (ii) minimize the following quantity:

$$\min_{\gamma} \int_{\Omega_0 \times \Omega_1} c(x, y) d\gamma(x, y),$$

for some *cost function* $c : \Omega_0 \times \Omega_1 \to \mathbb{R}^+$ that represents the cost of moving a unit of mass from x to y.

In the context of temporal change detection, the attention is directed towards identifying noteworthy deviations between consecutive days. As observed in empirical studies such as [59], the type and volume of traffic received by a Darknet are influenced by the IP space it monitors and its geographical location. Hence, this approach can also be utilized to gauge the degree of dissimilarity between two Darknets that monitor distinct dark IP spaces. In these scenarios, two cluster assignment matrices, M_0 and M_1, indicate the

clustering results for day-0 and day-1, correspondingly. These matrices are binary and have dimensions of $N \times K$, representing the cluster assignments for all N scanners, where $M_t \mathbf{1}_K = \mathbf{1}_N$ for $t \in 0, 1$. It is important to note that the number of scanners can fluctuate on different days; however, this variability does not compromise the overall applicability of the approach. The primary objective is to detect significant changes between the clustering outcomes M_0 and M_1 that would indicate changes in the Darknet structure from day-0 to day-1.Subsequently, the problem will be framed in terms of comparing two multivariate distributions, utilizing principles from optimal mass transport.

In the context of darknet clustering, the discrete version of the Kantorovich formulation is utilized. Here, the probability density functions I_0 and I_1 can be represented as $I_0 = \sum_{i=1}^{K} p_i \delta(x - x_i)$ and $I_1 = \sum_{j=1}^{K} q_j \delta(y - y_j)$, respectively, over the same space Ω, where $\delta(x)$ denotes the Dirac delta function. As a result, the optimal transport plan problem is transformed to:

$$\begin{aligned} K(I_0, I_1) &= \min_{\gamma} \sum_i \sum_j c(x_i, y_j)\gamma_{ij} \\ \text{s.t.} \sum_j \gamma_{ij} &= p_i, \sum_i \gamma_{ij} = q_j \\ \gamma_{ij} &\geq 0, i, j = 1 \ldots, K. \end{aligned} \tag{4.1}$$

Standard linear programming methods can be used to find solutions to this optimal transport plan problem. Moreover, when the cost function takes the form of $c(x, y) = |x - y|^p$, where $p \geq 1$, the optimal solution of Equation (4.1) defines a metric on the set of probability densities, $P(\Omega)$, which are supported on the space Ω. This metric is known as the *p-Wasserstein distance*, and it is defined as follows:

$$W_p(I_0, I_1) = \Big(\sum_i \sum_j |x_i - y_j|^p \gamma^*_{ij} \Big)^{\frac{1}{p}}, \tag{4.2}$$

where γ^* is the optimal transport plan for Equation (4.1).

The methodology involves the utilization of the 2-Wasserstein distance to assess the distributions I_0 and I_1, which are understood to encapsulate the clustering results M_0 and M_1. Here, the clustering assignment matrices M_u, for $u = 0, 1$, represent the outcomes for two consecutive days. Additionally, let X_0 and X_1 denote the $N \times P$ matrices that characterize the scanner features, as illustrated in Figure 3.2, corresponding to the two monitoring intervals. In order to determine the weights and Dirac locations for the discrete distributions I_0 and I_1, the following definitions are introduced:

$$D_u = M_u^\top \mathbf{1}_N \tag{4.3}$$
$$C_u = \Big(X_u^\top M_u\Big)\mathrm{diag}(D_u^{-1}), \quad u = 0, 1.$$

where D_u is a vector whose i-th entry represents the cluster size of the i-th cluster of scanners identified for day-u, and C_u is a matrix whose i-th row represents the clustering center of cluster i. Therefore, it is easy to determine the weights and Dirac locations for the discrete distributions $I_0 = \sum_{i=1}^{K} p_i \delta(x - x_i)$ and $I_1 = \sum_{j=1}^{K} q_j \delta(y - y_j)$. For instance, the weight p_i for cluster i on day 0 can be computed by normalizing the size of that cluster by the total number of scanners for that day, and the location x_i corresponds to the center of cluster i. Consequently, one can solve the minimization problem shown in to obtain the distance $W_2(I_0, I_1)$ and the optimal plan γ^*.

In Section 4.5, it is demonstrated that the distance $W_2(I_0, I_1)$, along with its corresponding optimal transport plan γ, can be effectively employed to (i) detect and (ii) interpret alterations in clustering between two consecutive monitoring windows. An alert indicating a change in clustering is triggered when the distance $W_2(I_0, I_1)$ is deemed "sufficiently large." However, it is noted that there exists no test statistic for the multivariate "goodness-of-fit" problem under investigation, in contrast to the univariate scenario. Consequently, the detection of anomalies is conducted by utilizing historical or empirical values of the $W_2(I_0, I_1)$ metric that can be collected. Upon the raising of an alert, the optimal transport plan γ is utilized to provide insights into the observed changes in clustering.

4.5 Evaluation

The evaluation of the proposed method utilized a comprehensive dataset comprising darknet traffic logs spanning the entirety of September 2016. This period witnessed the rapid proliferation of the Mirai botnet [36], a significant event characterized by a dramatic surge in infected Internet of Things (IoT) devices. Figure 4.1 illustrates this exponential growth in infections. Crucially, this escalation followed a discernible shift in Mirai's scanning strategy on September 6th, with the addition of TCP port 2323 to its target list. The subsequent exponential increase in compromised devices, commencing on September 14th, highlights a critical opportunity for timely intervention. The ability to detect such behavioral changes earlier, ideally by September 14th, could have enabled

preemptive mitigation strategies and potentially prevented subsequent attacks launched during the latter part of September 2016. This underscores the necessity for the proposed clustering and change-point detection framework, designed to address this critical gap in proactive security measures.

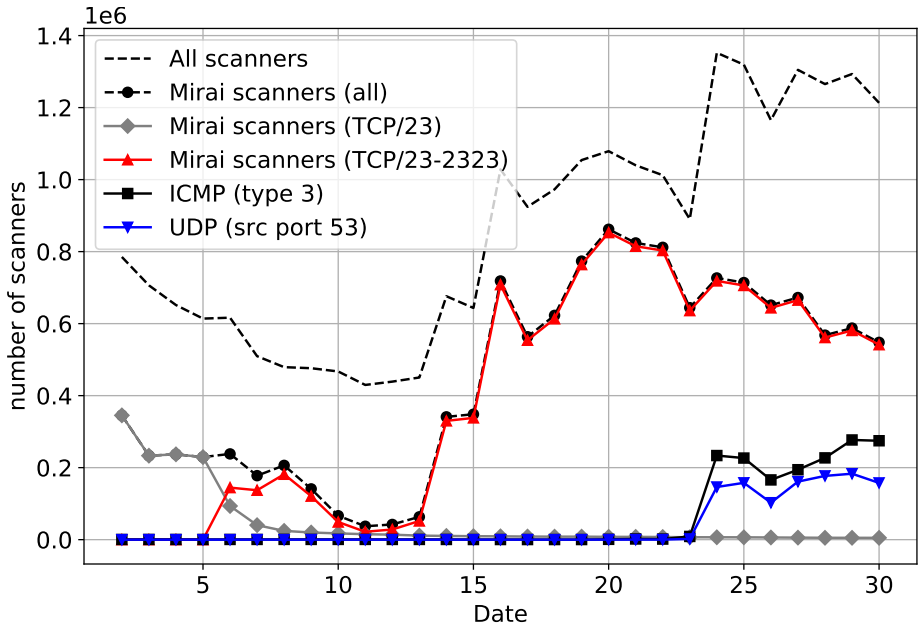


Figure 4.1: Evolution of the Mirai botnet depicted in Merit's Darknet scanning traffic for September 2016. The graph shows the addition of TCP/2323 in the set of scanned ports, with a minimum of 50 packets emitted daily by the scanners.

The dataset's scale and characteristics are detailed in Table 4.1. Over the month, the darknet traffic included around 35 million unique source IP addresses, marking a significant increase compared to previous research like DANTE [66] and DarkVec [67], with two orders of magnitude more unique sources. Additionally, the dataset's completeness is impressive, as every possible port destination (0-65535) has at least one scan attempt.The comprehensive nature of the dataset provides a robust foundation for evaluating our proposed method under realistic conditions. The high volume of unique source IPs and extensive port coverage enhance its richness and suitability for detecting subtle shifts in botnet activity within a complex, high-volume network environment. Further analysis will leverage these characteristics to rigorously assess the framework's performance against established metrics of accuracy and detection latency.

Table 4.1 reveals that Mirai-related ports, predominantly TCP/23 and TCP/2323, constitute a substantial 65% of the total observed traffic. This high proportion underscores the significance of Mirai in the Darknet landscape during the observation period.

Table 4.1: Basic statistics for our Darknet datasets.

Dates	Darknet Size	Sources	Packets	Ports	Top-3 ports		
					Port	Traffic (%)	Sources
[2016-09-02, 2016-09-30]	/10	35M	49B	65536	23 80 2323	60.34 13.55 4.00	20.5M 963K 13.5M
2016-09-14	/10	1.8M	1.5B	65536	23 2323 80	53.30 11.39 6.83	808K 527K 96K
2016-09-24	/10	3.3M	1.4B	65536	23 2323 80	69.45 7.00 3.73	1.8M 1.3M 84K
2022-02-20	/13	845K	3.1B	65536	6379 23 22	6.67 5.10 2.17	2.5K 122K 10.4K

To effectively characterize Darknet activity, it is crucial to address the inherent noise within the dataset. This noise stems from various sources including misconfigurations and randomly spoofed source IP addresses. A common mitigation strategy, adopted here and in previous studies [53, 67, 85], involves filtering out scanners transmitting fewer than 50 packets. This threshold balances the removal of spurious data with the preservation of meaningful scanning activity. The rationale behind this threshold is twofold: firstly, it reduces noise caused by transient or erroneous connections, thereby improving the accuracy of subsequent analyses; secondly, certain features used to characterize Darknet probes, such as average packet inter-arrival times, require a minimum number of observed packets for reliable computation.

Figure 4.2 depicts the cumulative count of unique source IP addresses observed over time, both with and without the application of the 50-packet filter. The figure clearly shows the impact of filtering, demonstrating a more focused representation of sustained scanning activity. Furthermore, the figure illustrates the temporal growth of the Mirai botnet, evident in the increasing number of unique source IPs. The persistent presence of a subset of scanners, defined as those active throughout the entire month, is also highlighted, providing insights into the sustained nature of some scanning campaigns. This persistent activity is a crucial factor for identifying and understanding the long-term impact of these attacks.

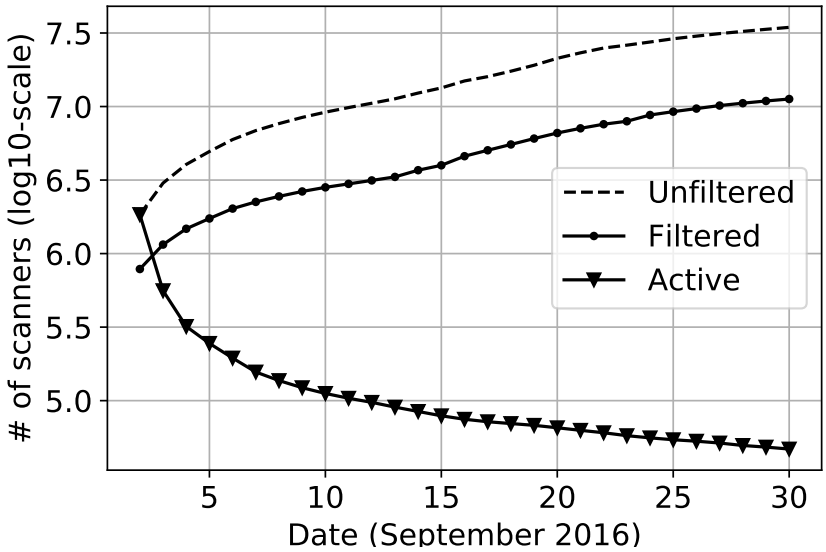


Figure 4.2: Evolution of scanning activity over time.

4.5.1 Evaluation Challenges

The rigorous evaluation of temporal change detection techniques within the realm of cybersecurity, specifically regarding darknet traffic analysis, is recognized as a significant challenge due to the intrinsic difficulty associated with acquiring reliable ground truth data. To address this limitation, two principal methodologies have been employed: manual labeling and synthetic data generation.

Manual labeling, although potentially yielding highly accurate ground truth, is characterized by a heavy reliance on subjective expert judgment. This process typically involves visual inspection, statistical analysis, and heuristic approaches [86]. It is acknowledged that this methodology is inherently labor-intensive, costly, and prone to inter-rater reliability issues stemming from individual biases, which compromises the objectivity and consistency of the ground truth labels.

To mitigate the aforementioned limitations, researchers often resort to the generation of synthetic data. This approach entails the creation of simulated datasets that model the underlying processes being monitored, with controlled changes introduced, such as gradual drifts, abrupt shifts, and periodic fluctuations at defined time points. The advantage of this method resides in the inherent knowledge of the ground truth, which facilitates an unbiased evaluation of the accuracy of change detection algorithms and enables robust analyses across various techniques. Moreover, the controlled environment permits a systematic investigation into the sensitivity of algorithms to different parameters and the exploration of their performance under diverse change scenarios.

In this study, the synthetic data generation approach is leveraged through the employment of a Bayesian network to model the complex probabilistic dependencies inherent in darknet traffic features. This Bayesian network effectively captures the intricate relationships among various traffic characteristics, thereby providing a more realistic representation of the darknet traffic generation process compared to simpler models. Synthetic datasets are generated by sampling from the learned joint probability distribution of these traffic features. It is crucial to acknowledge that, despite the sophistication of the modeling approach, simulated data may not perfectly replicate the full complexity and inherent variability of real-world darknet traffic. The assumptions embedded within the Bayesian network, while carefully considered, may not encompass all the nuances of the actual data-generating process. Consequently, to enhance the generalizability and robustness of the findings, the performance of the temporal change detection techniques is further validated using a separate corpus of real-world darknet traffic data exhibiting known temporal changes. This two-pronged evaluation strategy, which encompasses both synthetic and real-world data, ensures a more comprehensive and rigorous assessment of the proposed methods.

4.5.2 Evaluation on Synthetic Data

This dissertation investigates the efficacy of novel clustering and temporal change detection algorithms within the domain of network traffic analysis. A comprehensive experimental design is employed to rigorously evaluate these algorithms, leveraging both synthetic and real-world datasets. The selection of synthetic data generation methods is recognized as crucial, given its direct impact on the generalizability and robustness of the evaluation results. A range of techniques exists for generating synthetic data, each with its own strengths and limitations.

Statistical approaches, such as bootstrapping, Gaussian mixture models, and Markov chain Monte Carlo methods, offer computationally efficient solutions; however, they may struggle to capture the complex, high-dimensional dependencies inherent in real-world network traffic. Rule-based methods, while providing control over specific features, often lack the necessary flexibility required to accurately represent the intricate relationships within the data. Conversely, deep learning models, including Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), are capable of generating highly realistic data but necessitate substantial computational resources, extensive training, and careful hyperparameter tuning. Furthermore, their "black box" nature severely limits interpretability and understanding of underlying processes.

Given the need to model complex dependencies and represent uncertainty within network flow data, Bayesian networks are utilized for synthetic data generation in this study. Bayesian networks, as probabilistic graphical models that represent probabilistic relationships between variables, are particularly well-suited for this purpose. Their ability to explicitly model causal relationships and incorporate prior knowledge provides a principled framework for generating realistic synthetic data while maintaining control over specific features. Additionally, the inherent capacity to sample from the joint probability distribution facilitates the simulation of various "what-if" scenarios, offering granular control over data characteristics. The efficacy of Bayesian networks in modeling network flow traffic has been previously demonstrated [80], and the availability of a substantial dataset in this study mitigates the limitations encountered previously due to insufficient training data.

The performance evaluation of the proposed clustering and temporal change detection algorithms proceeds in two phases. Initially, the algorithms are evaluated on synthetic datasets generated using the aforementioned Bayesian network approach. The synthetic data incorporates known change points, allowing for a precise assessment of model fit and accuracy. Although this approach is inherently constrained by the predefined structure of the synthetic data, a diverse range of "what-if" scenarios is rigorously introduced to probe the robustness of the algorithms under varying conditions, thereby mitigating this limitation. These scenarios explore different magnitudes, frequencies, and types of changes in the network traffic patterns.

Subsequently, thorough validation is conducted using multiple real-world datasets that exhibit diverse change scenarios observed in actual network traffic. This two-pronged approach—encompassing controlled experiments with synthetic data and validation with real-world datasets—offers a robust and comprehensive evaluation of the proposed algorithms, ensuring both internal and external validity. The results from both phases are meticulously analyzed to establish the efficacy and limitations of the proposed methods under realistic conditions.

4.5.2.1 Synthetic Data Generation

To learn the Bayesian network, the *hill-climbing* algorithm implemented in R's `bnlearn` package [87] is employed. The training of this Bayesian network is conducted exclusively on the numerical features, specifically those categorized under "Network Activity Metrics" and "Scanning strategy", with the exception of the *destination strategy*, *IPID strategy*, and *IPID options*. The purpose of this training is to capture the causal relationships

among the selected features. The resulting network is structured as a directed acyclic graph (DAG), where nodes represent the features and directed edges between node pairs signify the conditional dependencies that exist between them.

Let $G = (V, E)$ denote a *directed acyclic graph*, with vertices $V = \{X_1, X_2, \ldots, X_n\}$ representing the numerical features, and directed edges E indicating the conditional dependencies among the features. The joint probability distribution of the features, denoted as $P(X_1, X_2, \ldots, X_n)$, can be articulated using the chain rule of probability as:

$$\mathbb{P}(X_1, \ldots, X_n) = \prod_{i=1}^{n} \mathbb{P}(X_i | \text{parents}(X_i)) \tag{4.4}$$

where parents(X_i) denotes the set of parents of X_i in G. It can be shown that for every variable in the network X_i, we can have:

$$\mathbb{P}(X_i | X_{i-1}, \ldots, X_1) = \mathbb{P}(X_i | \text{parents}(X_i)) \tag{4.5}$$

A topological order of the nodes in the Bayesian network can ensure that the relationship is met [88]. Once the joint distribution is specified, a Monte Carlo sampling algorithm generates data points for each input parameter based on the probability distribution [88]. For the Monte Carlo method, we consider the variables $X_1, \ldots, X_n$ to be Gaussian random variables with a joint distribution of $\mathcal{N}(\mu, \Sigma)$. To do this, we make use of the conditional distribution relationships that apply to multivariate Gaussian random variables. The values of the parameters μ and Σ are estimated using the same real Darknet dataset we use to learn the Bayes net.

After generating numerical features using the Monte Carlo approach, we include the "set of ports scanned" feature to combine categorical and numerical features in each synthetic data point. To create K separate clusters, we space the values of the root nodes in the Bayes network accordingly. The Bayesian network in Figure 4.3 depicts the conditional dependencies observed among the numerical darknet features. In this generative model, the input values for the root nodes "NumPorts" and "AvgInterArrival" are defined based on the temporal change scenario under examination. The model then generates data points based on the joint probability distribution it has learned.

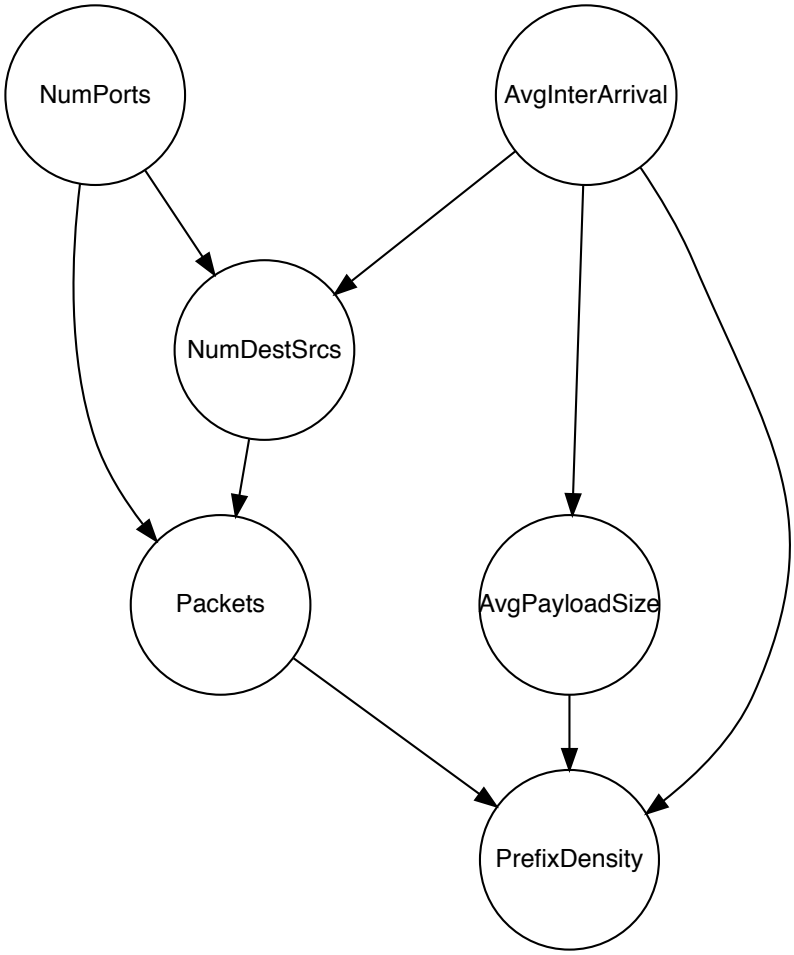


Figure 4.3: Bayesian network graph depicting the conditional dependencies between various numerical darknet features. The feature pointed by the arrowhead is dependent on the feature at the arrow's tail.

4.5.2.2 Sensitivity Analysis

This section investigates the sensitivity of the 2-Wasserstein distance, $W_2(I_0, I_1)$, a critical component of our temporal change detection framework, to variations in input distributions. Understanding this sensitivity is crucial for evaluating the robustness and reliability of our framework in real-world network traffic analysis scenarios. The approach employed utilizes Monte Carlo simulations to quantify the impact of perturbations in the input distributions on the calculated Wasserstein distance. This methodology involves

generating synthetic network traffic data reflecting realistic change scenarios, and then systematically analyzing the resulting changes in $W_2(I_0, I_1)$.

Two prevalent scenarios encountered in network security are specifically focused upon: (1) gradual variations in the scanning traffic volume, exemplified by the increasing activity of a botnet due to the growing number of compromised devices and (2) incremental modifications in scanning strategies, reflecting the evolving attack methodologies, such as the broadening of targeted ports or the exploitation of newly identified vulnerabilities within an attacker's toolkit, including alterations to exploit kits.

Synthetic data representing these scenarios is generated using the Bayesian network detailed in Section 4.5.2.1. This generative model allows for the controlled introduction of subtle changes in the underlying distributions, enabling a precise assessment of the W_2 metric's sensitivity. The Monte Carlo simulations involve repeated sampling from perturbed distributions, calculation of the resulting W_2 distances, and subsequent statistical analysis to quantify the relationship between input perturbations and the output distance. This provides a robust measure of the framework's resilience to noise and minor variations in the observed network traffic. The results of this sensitivity analysis are presented in Figure 4.4.

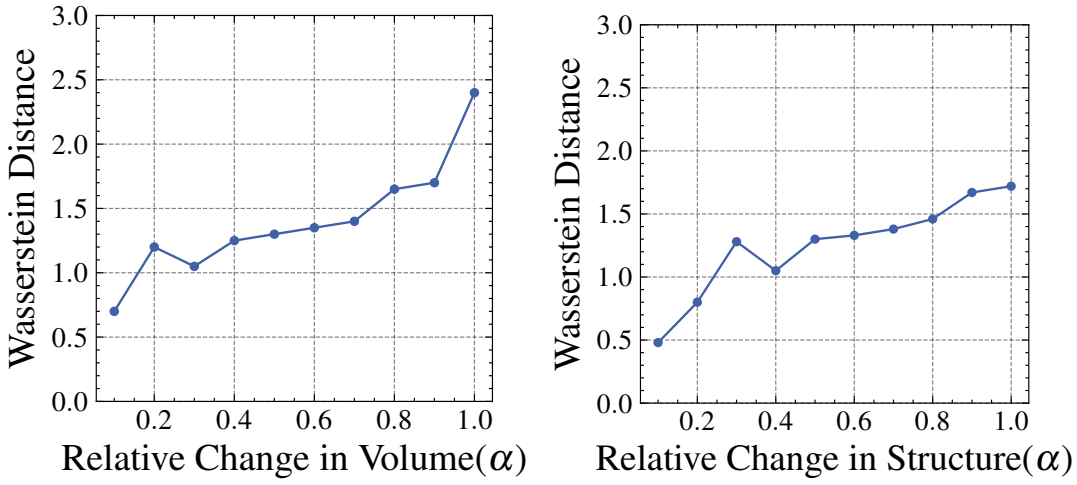


Figure 4.4: The Wasserstein distance exhibits considerable variance in response to alterations in both volume (left) and structure (right) of the input distributions.

4.5.3 Evaluation on Real World Darknet Data

To rigorously evaluate the efficacy of the proposed framework in discerning distributional shifts within network traffic data, empirical analyses were conducted using two distinct real-world datasets. The selection of these datasets was aimed at capturing a variety of network conditions and adversarial scenarios. An initial validation was performed using a dataset that encompassed network traffic logs collected over a one-month period in September 2016, coinciding with a significant surge in activity associated with the Mirai botnet [36]. This period, as visually depicted in Figure 4.1, served as a challenging testbed due to the inherent complexity and scale of the Mirai botnet's network operations. The comprehensive application of the framework to this dataset acted as a critical benchmark, establishing the baseline performance of the proposed clustering methodology under conditions characterized by high network traffic volume and sophisticated botnet activity. This extensive evaluation ensured the reliability and overall performance of the entire pipeline, which included data preprocessing, feature extraction, clustering, and anomaly detection.

Subsequently, to further assess the robustness and generalizability of the clustering technique, a second dataset was employed, comprising network traffic data collected on February 20, 2022. This dataset, which was temporally distinct from the first, enabled the evaluation of the method's performance under potentially different network conditions and attack vectors. The contrast between these two datasets—one characterized by large-scale, known Mirai botnet activity and the other representing an uncharacterized snapshot of network traffic—provided a robust assessment of the proposed framework's adaptability to diverse network scenarios and its resilience to unforeseen events.

This two-pronged empirical validation approach significantly strengthened confidence in the generalizability of the framework's conclusions beyond the specific context of the Mirai botnet. The choice of these specific dates was influenced by the availability of suitably labeled and high-quality data, along with the contextual relevance necessary for evaluating the method's performance against varying threat landscapes. The results of this evaluation, demonstrate the consistent effectiveness of the proposed approach across a range of conditions, thereby reinforcing its utility in addressing the evolving challenges within network security.

4.5.3.1 Mirai Onset: September 2016

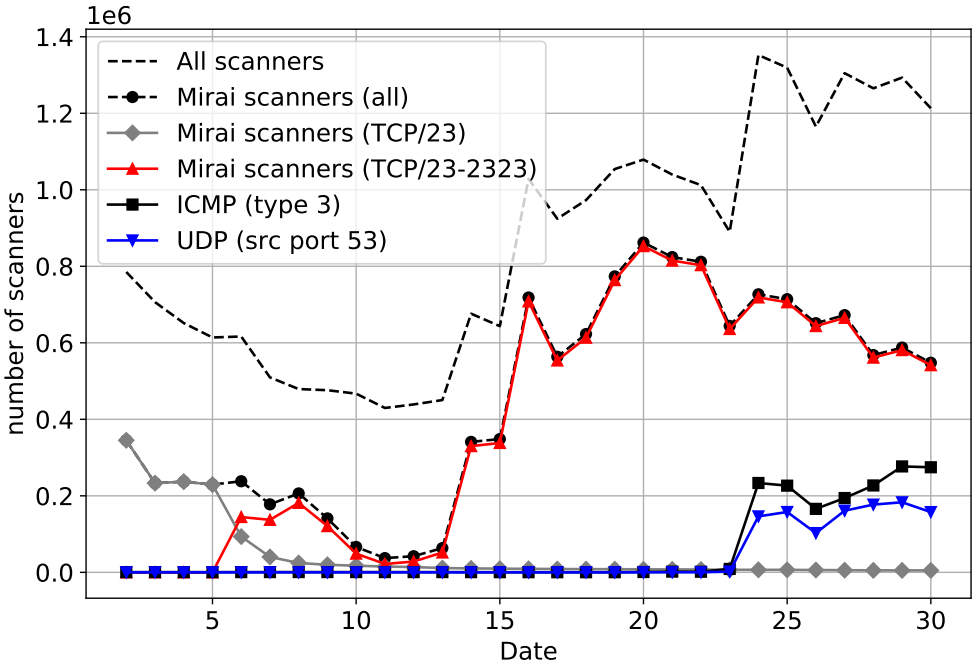
The Mirai botnet, first identified in August 2016, launched large-scale Distributed Denial-of-Service (DDoS) attacks in September of the same year. This period represents a critical juncture in the malware's evolution and has been extensively studied. To rigorously evaluate the proposed framework, we employed a dataset encompassing network traffic captured from September 2nd, 2016, onwards. This selection allows for a comprehensive analysis of the botnet's activity during a period of intense operational activity and known evolutionary changes. The dataset comprises raw network flow data which was pre-processed to extract relevant scanner features. These features were subsequently converted into low-dimensional embeddings using a pre-trained autoencoder. A daily analysis was performed. Each day's scanner embeddings were clustered into 200 clusters via k-means clustering to manage the inherent dimensionality and heterogeneity of the botnet's activity. The Wasserstein distance metric and corresponding optimal transport plan were then calculated between consecutive days. This approach provides a robust measure of the temporal evolution of the botnet's scanning behavior, allowing us to quantify changes in the attack landscape over time. The choice of 200 clusters was informed by an elbow method analysis of the clustering performance.

The lower graph in Figure 4.5 illustrates the time-series of 2-Wasserstein distances for September 2016. At a significance level of 5%, two distinct change points can be discerned. The first change point was observed on September $14th$ (with a corresponding p-value of 0.036), while the second change point was identified on September $24th$ (with a p-value of 0). The calculation of p-values was based on the complete set of estimated Wasserstein distances derived for the entirety of the month[1].

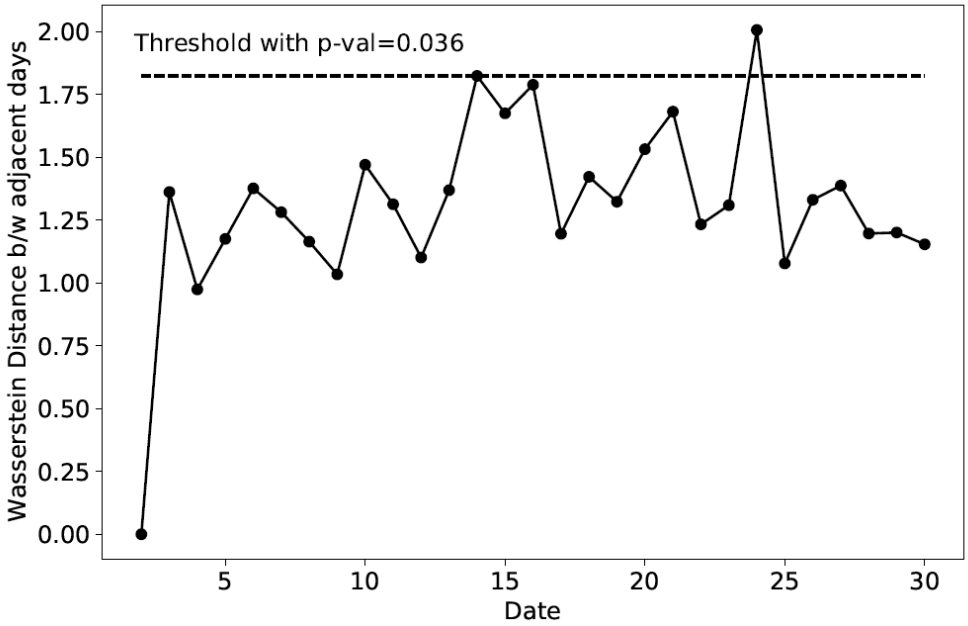
Following the identification of the change points, the ensuing step entails employing the optimal transport plan γ^* to *interpret* the detected change-points. The optimal transport plan from cluster distribution A to B can be represented as a weighted directed graph $G = (V, E)$ with nodes $V := \{A_u\} \cup \{B_u\}, u = 1, \ldots, K$, where node A_u corresponds to cluster-u in day-i and B_u to cluster-u in day-$i+1$, respectively. An edge $(u, v) \in E$ exists if and only if $\gamma^*_{uv} > 0$, signifying that some mass has been transferred from cluster-u of day-i to cluster-v of day-$i+1$. The weight w_{uv} assigned to each edge $(u, v) \in E$ is given by γ^*_{uv}.

Figure 4.6 showcases the graph resulting from the optimal transport plan γ^* for the clustering outcomes on September 13 and September 14. Upon initial observation, the

[1]In real-world applications of our system, historical Wasserstein values may be utilized, including those obtained from the previous month



(a) The graph shows the addition of TCP/2323 in the set of scanned ports, with a minimum of 50 packets emitted daily by the scanners.



(b) This graph represents the detection of temporal changes in the Darknet using the Wasserstein distance.

Figure 4.5: Expansion of the Mirai botnet depicted in Merit's Darknet scanning traffic for September 2016 and its detection using Wasserstein distance.

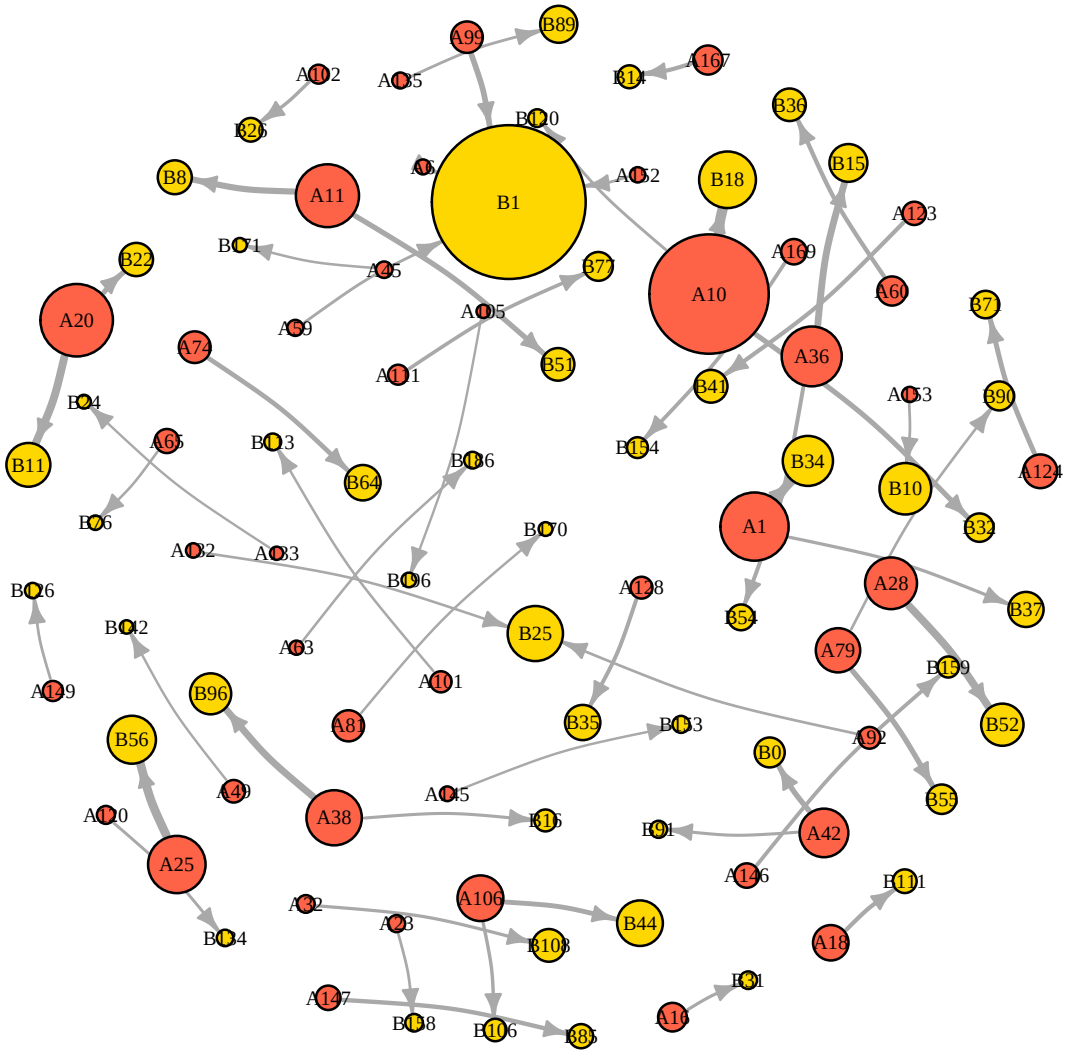


Figure 4.6: Optimal transport plans for Sept. 13–14. Only edges with $\gamma^*_{uv} \geq 0.01$ are shown.

.

graph reveals the maximum mass transferred from cluster A_{10}, the largest cluster on September 13, to cluster B_{18} on September 14. Both clusters, A_{10} and B_{18}, correspond to Mirai compromised bots that scan port TCP/23. The relatively smaller size of B_{18} in comparison to A_{10} indicates a declining trend in the number of Mirai-related scanners targeting port TCP/23. This conclusion is further supported by the second-largest mass transfer from cluster A_1 to B_{34}, where B_{34} represents the introduction of port TCP/2323 in the set of ports scanned by Mirai. Additional insights can be obtained by examining cluster pairs such as (A_{25}, B_{56}), (A_{20}, B_{11}), (A_{28}, B_{52}), (A_{38}, B_{96}), and other such pairs, which are not included here due to space constraints. Upon examination of the upper graph in Figure 4.5, it becomes apparent that the change in the clustering outcomes

between September 13 and 14 can be ascribed to the evolving tactics employed by the Mirai botnet. However, it is noteworthy that detecting such changes without the aid of an automated methodology, as proposed in this study, would necessitate monitoring an unwieldy amount of time series, such as the scanning traffic directed towards all ports, rendering it practically infeasible.

The most significant change during this period of Mirai emergence, however, was identified during September 23–24, as depicted in the right panel of Figure 4.5. Intriguingly, this change coincided with a remarkable surge in the volume of darknet traffic related to UDP scanning and ICMP messages with Type 3 (Destination Unreachable). Upon closer examination of the traffic, it was observed that the majority of the UDP packets had `src port 53` and the ICMP messages contained the message `destination port 53 unreachable`. Further analysis of the payloads of these packets led to the identification of a significant amount of malicious DNS scanning activity, captured in darknet as "DNS backscatter" [60]. The UDP and ICMP packets contain DNS A-record queries under the domain `xy808.com` with randomly generated subdomains, which is a well-known strategy employed by scanners to identify open DNS resolvers while masking their identity. These lists of open DNS resolvers are often utilized in volumetric, reflection, and amplification DDoS attacks, as documented in [89]. Considering the events of September 25th and October 21st, 2016, when significant DDoS attacks were launched against Krebs on Security and Dyn, it is reasonable to assume that the actors behind the Mirai botnet were responsible for the intense DNS scanning activities observed during that time. The link between these attacks and the scanning activities suggests a deliberate effort to exploit vulnerabilities in network infrastructure.

Table 4.2 provides a summary of the optimal transport plan graph G for September 23–24, emphasizing the top 6 pairs of clusters that exhibit the highest amount of mass transfer. This table is essential for understanding the dynamics of mass movement across different clusters during this timeframe. Of particular interest is the last row of Table 4.2, which reflects a significant transfer of mass from cluster A_{47} on September 23 to cluster B_{24} on September 24. This transfer is notable as it indicates the formation of a completely new cluster, suggesting a shift in the overall structure of the clustering distribution. In contrast to other clusters identified in the clustering distribution from September 23, cluster B_{24} displays a distinctly different scanning profile. This profile is characterized by activities related to ICMP (type 3), which points to a unique behavior not seen in the other clusters. Moreover, this difference is further underscored by a Jaccard similarity score of 0, indicating that cluster B_{24} is dissimilar to all other clusters.

Table 4.2: Interpretation of clustering changes between September 23 and September 24, 2016. Notice that the last row indicates the formation of a new large cluster (cluster 24), associated with a DDoS attack.

Day	Label	Mass	Jaccard	Traffic	Freq.	Ports	Freq.
23	13			TCP-SYN	22208	23-2323	22099
24	63	0.025	0.14	TCP-SYN	37520	23-2323	37322
23	9			TCP-SYN	20539	23-2323	20430
24	60	0.023	0.16	TCP-SYN	31195	23-2323	29094
23	28			TCP-SYN	24141	23-2323	19273
24	25	0.022	0.12	TCP-SYN	29387	23-2323	21269
23	81			TCP-SYN	31094	23-2323	31028
24	1	0.021	0.18	TCP-SYN	32536	23-2323	32437
23	11			TCP-SYN	23787	23-2323	21545
24	29	0.021	0.11	TCP-SYN	28583	23-2323	26336
23	47			TCP-SYN	19702	23-2323	19592
24	24	0.017	0.00	ICMP (type 3)	23146	0	23204

Figure 4.7 illustrates the in-degrees of graph G derived from the optimal transport plan of September 23–24. The three panels depict the pruned edges where $\gamma^*_{uv} < \tau$ with τ being a threshold value from the set $\{5 \times 10^{-4}, 0.001\}$. Notably, cluster B_{123} has the highest in-degree in all three cases, indicating its significance. The high mass transfers from various clusters of the previous day to cluster B_{123} through the optimal transport plan suggest that it is a novel cluster. This assertion is corroborated by the cluster members, which are associated with UDP messages with `src port 53`. As indicated in Figure 4.5, this activity started on September 24th.

4.5.3.2 Cluster Inspection: February 2022

The efficacy of threat intelligence analysis can be significantly enhanced through the application of clustering algorithms to identify emergent patterns and characterize existing threats within the complex threat landscape. This section presents a case study employing a novel clustering approach on a specific dataset, illustrating its utility in revealing actionable insights. The analysis focuses on data collected on February 20th, 2022, representing the most recent available dataset at the time of this research. This date was chosen due to its representative nature of the observed threat activity.

Our methodology involved a multi-stage process. Initially, a comprehensive dataset comprising approximately 845,000 network scanners was obtained from Merit's darknet observation system. Subsequently, a rigorous filtering process was implemented to

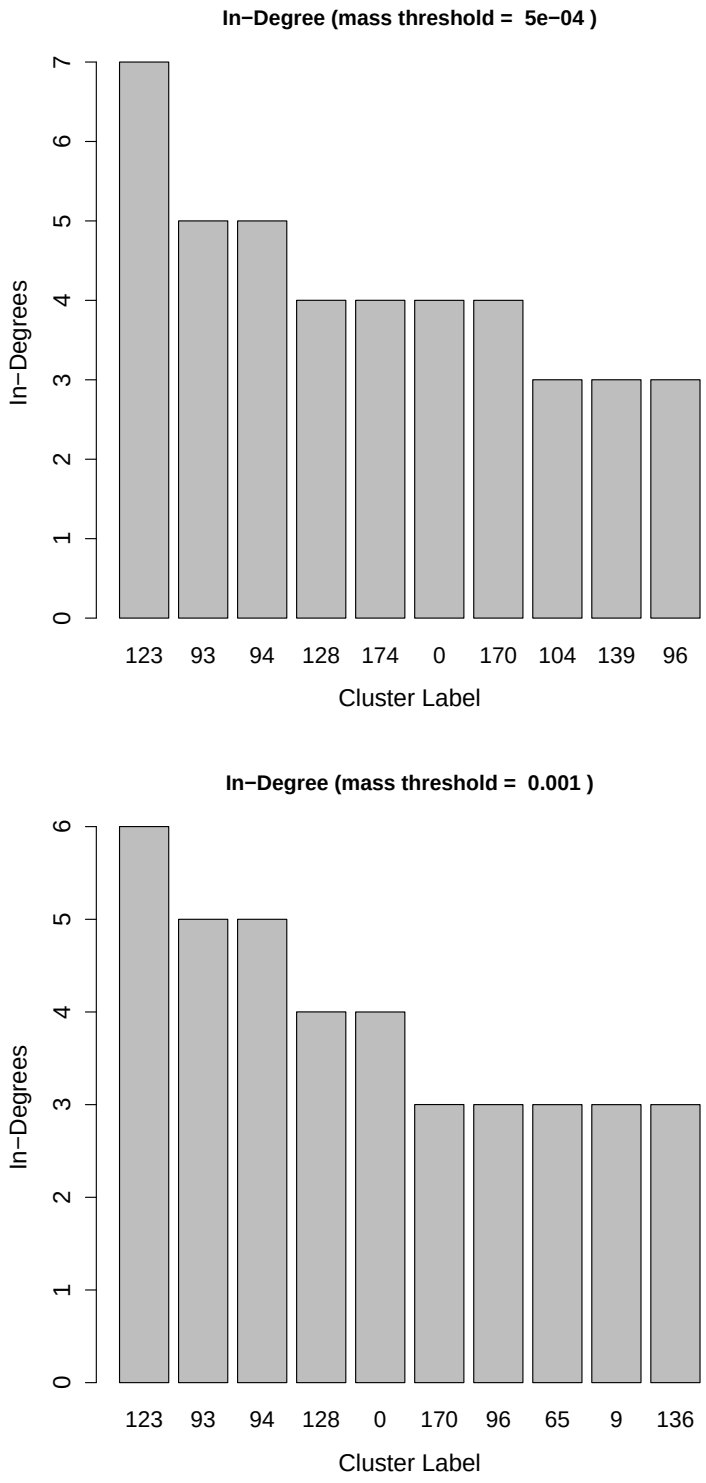


Figure 4.7: In-degree distributions of the graph induced by the optimal plan γ^* for Sept. 23–24.

eliminate low-volume senders, resulting in a refined dataset of 223,909 scanners. This reduction significantly improved the signal-to-noise ratio, facilitating more accurate clustering and subsequent analysis. The characteristics of these remaining scanners were then subjected to an unsupervised clustering algorithm (detailed in Section 3.2), yielding distinct groupings based on shared behavioral attributes. The resulting clusters are summarized in Table 4.3, categorized according to prominent characteristics. These categories provide a high-level overview of the prevalent scanning activities observed on the chosen date, allowing for a refined understanding of the threat landscape's composition on February 20th, 2022. Further investigation into the individual clusters reveals granular details about the specific threat actors and their operational methods.

Table 4.3: Cluster Inspection (2022-02-20).

Description	**# of Clusters**	**# of Senders**
Mirai-related	70	108,912
Unknown	67	76,525
SMB	20	23,700
Heavy Scanners	19	2,377
ICMP scanning	5	2,619
Ack Scanners	4	795
SSH scanning	4	2,635
censys.io	3	147
TCP/3389 (RDP)	2	1,482
UDP/5353	2	3,212
Backscatter (DDoS)	2	815
TCP/6379 (Redis)	1	437
Normshield	1	253
TOTAL	**200**	**223,909**

Our analysis revealed the existence of 70 clusters comprised of a total of 108,912 scanners that are classified as "Mirai-related" based on their targeted destination ports and traffic type, which is TCP-SYN. It is important to note that not all of them exhibit the characteristic *Mirai fingerprint*, which involves setting the scanned destination address equal to the TCP initial sequence number, as reported in [36]. This suggests the presence of multiple Mirai variants, which is also evidenced by the existence of clusters that scan the telnet port "23" along with other various combinations of ports such as "23", "23-2323", "23-80-8080" and "5555". More complex sets of ports such as "23-80-2323-5555-8080-8081-8181-8443-37215-49152-52869-60001" have also been observed, which

have been linked to recent strains of Mirai, such as Mozi [90]. Substantial majority of these clusters appear with TTL fields typical of Linux/Unix systems, which strongly suggests that these clusters are predominantly composed of IoT/embedded devices that have been compromised [91].

This section analyzes the diverse characteristics of detected Darknet scanning clusters, revealing distinct patterns indicative of various threat actors and scanning methodologies. Our analysis categorizes these clusters into several distinct groups based on observed behaviors, target ports, and inferred operating systems.

The first group comprises clusters exhibiting anomalous scanning behavior not readily attributable to known malware families or threat actors, hereafter designated as "Unknown." This group predominantly utilizes UDP traffic targeting high-numbered, dynamically changing ports. Analysis of the Time-To-Live (TTL) field suggests a heterogeneous operating system distribution, encompassing both Windows and Linux/Unix systems. Geospatial analysis indicates a significant concentration of these scanners within China.

A second notable group consists of 20 clusters exhibiting a strong association with TCP port 445 scanning, characteristic of the Server Message Block (SMB) protocol. Exploitation of SMB vulnerabilities is a well-documented tactic for various ransomware variants, including, but not limited to, WannaCry [92, 93]. The constituent machines of these clusters are predominantly Windows-based.

Further analysis identified a substantial number of "heavy scanner" clusters, encompassing both benign (e.g., Censys [57] and Shodan [94]) and potentially malicious actors. Four clusters consist almost exclusively of acknowledged scanners – those originating from known research institutions and other entities deemed non-hostile [95]. An additional four clusters, three originating from Censys and one from Normshield [96], display benign scanning behavior from IP addresses not yet included in the acknowledged scanner list [95]. Several heavy scanner clusters demonstrated notable anomalous behaviors, including: high-speed scanning (five clusters with mean inter-packet arrival times $<$ 10ms); exhaustive or near-exhaustive scans of all monitored Darknet IPs (ten clusters); near-exhaustive port scans (two clusters, scanning nearly all 2^{16} ports); substantial UDP payload transmission to 16 distinct ports (one cluster); and extensive SIP scanning (two clusters).

Our investigation also uncovered a significant cluster (437 scanners) targeting TCP port 6379 (Redis). Data from Table 4.1 reveals TCP/6379 as the most frequently scanned port (by packet count) on 2022-02-20. Cluster analysis revealed a high degree of

homogeneity within this group, characterized by high scanning frequency, near-exhaustive scanning of monitored Darknet IPs, predominantly Linux/Unix-based systems, and a geographical concentration in China. Additionally, we identified two clusters engaging in TCP/3389 (RDP) scanning, two targeting UDP/5353 (DNS), and two exhibiting "backscatter" activity consistent with spoofed-based DDoS attacks [97].

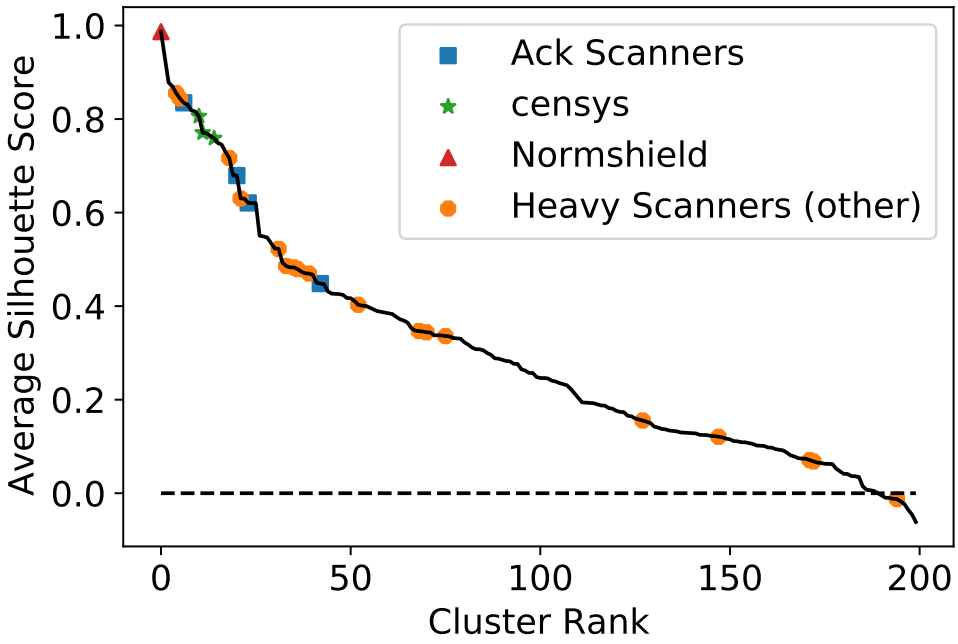


Figure 4.8: Average silhouette score for all clusters (2022-02-20).

The efficacy of our clustering algorithm is evaluated using the silhouette score, a metric quantifying the similarity of an object to its own cluster compared to other clusters. This metric ranges from -1 to 1, with higher values indicating better-defined, more compact clusters. A score of 1 represents perfect separation, while -1 signifies misclassification. Figure 4.8 presents the average silhouette scores for each cluster identified within the February 20th, 2022 dataset. We selected representative clusters exhibiting similar behavioral patterns – for example, clusters exhibiting consistent packet counts or targeting a similar number of ports – to demonstrate the algorithm's performance. This selection provides robust examples for evaluating the clustering outcome.

The majority of these selected clusters exhibit high silhouette scores (≥ 0.33), demonstrating a clear separation between distinct scanning behaviors. These clusters include four identified as "Acknowledged Scanners," three associated with Censys activity, and a single cluster attributed to Normshield. Furthermore, 18 clusters categorized as "Heavy Scanners" were similarly annotated (excluding singleton clusters corresponding to

NETSCOUT's research scanner, for which a silhouette score is undefined). While most clusters exhibit strong separation, some display lower scores due to inherent variability in certain features, such as Time-To-Live (TTL) values. For instance, cluster 162, with a silhouette score of -0.01, represents significant scanning activity targeting nearly all Darknet IPs, averaging 5,753 unique ports per scan. The relatively low score for this cluster is attributable to the diversity in its feature space, particularly the TTL values. This observation highlights a potential limitation of the algorithm, where nuanced variations in less prominent features can impact the silhouette score. However, even clusters with scores approaching zero retain analytical value. Cluster 162, for example, provides critical intelligence on a high-intensity scanning campaign targeting the Darknet. Should further refinement be desired, hierarchical clustering techniques could be employed to subdivide clusters, offering a pathway to enhance granularity and precision in the analysis.

4.6 Discussion

This chapter presents a novel framework for threat detection that leverages network telescope data. The framework has demonstrated significant promise in assisting security analysts in the development of accurate and efficient threat alert systems with low false positive rates. The integration of autoencoders for representation learning, along with the k-means algorithm for clustering, has been identified as a robust approach to analyzing high-velocity, high-dimensional network scanning data. Preliminary experiments indicated that this combination yielded superior performance compared to alternative methods; however, the inherent limitations of these techniques necessitate further investigation.

The autoencoder architecture, while effective in dimensionality reduction, is characterized by a task-agnostic nature that can limit its application in specialized contexts. Minimizing the reconstruction error between input and output does not inherently guarantee an optimal representation specifically tailored for threat detection purposes. As a result, the learned embeddings may not fully capture crucial domain-specific knowledge that is pertinent to identifying and distinguishing malicious scanning activity. This limitation suggests that further exploration of alternative, domain-specific embedding techniques is warranted in future research endeavors.

Furthermore, the k-means algorithm requires the pre-specification of the number of clusters, a parameter that was determined empirically in the current study through exhaustive experimentation. This reliance on pre-defined cluster numbers introduces potential bias and limits the algorithm's adaptability to datasets with varying character-

istics. Therefore, investigating alternative clustering algorithms, such as density-based methods that are less sensitive to the number of clusters, represents a crucial area for future development.

Despite these limitations, the efficacy of the proposed framework in handling challenges posed by high-speed data streams and the high dimensionality of network scanning features has been demonstrated. The framework successfully identifies clusters of scanners, thereby facilitating improved threat detection and risk assessment. However, it was observed that the resulting clusters may exhibit heterogeneity, particularly concerning the intensity and target port range of scanners within a single cluster. For instance, high-intensity scanners targeting a limited set of ports may be grouped with lower-intensity scanners exhibiting similar port scanning behavior. This necessitates the refinement of the clustering methodology.

A multi-pass clustering approach could be adopted, which would iteratively refine cluster assignments based on varying parameters or feature subsets while considering cluster size and feature entropy (e.g., entropy of scanned ports). Such refinements could significantly enhance the precision and granularity of the clustering results, leading to more meaningful clusters and enabling more effective threat identification and subsequent response actions.

Beyond enhanced threat detection, the clusters generated by the framework offer significant potential for broader security applications. The identification of high-intensity scanners provides valuable enrichment to existing scanning data, which is often deficient in detecting sophisticated attacks. This enriched dataset significantly improves risk assessments, as the identified target ports reveal potential vulnerabilities that adversaries may exploit. Consequently, enterprise security teams can proactively revise their risk profiles and develop more effective mitigation strategies.

Additionally, the output of the framework can facilitate the filtering and prioritization of Indicators of Compromise (IoCs) generated by Intrusion Detection Systems (IDS). This prioritization enhances the efficiency and effectiveness of incident response by focusing on the most critical alerts, thereby demonstrating the framework's versatility and practical impact beyond its primary threat detection function. Future research will explore the integration of these enhancements into a unified, automated system.

Future work could also investigate the characteristics of persistent scanners to gain a better understanding of their operational patterns and targets. This exploration would contribute to a deeper comprehension of the evolving threat landscape, further informing the development of adaptive security measures.

Chapter 5
IP Threat Intelligence Enhancement

5.1 Background

The efficacy of threat detection is significantly enhanced through the deployment of large-scale network telescopes, which provide extensive coverage of malicious internet activity. However, the identification of threats represents only the initial phase of effective cybersecurity. A crucial subsequent step involves accurately determining the motivations of threat actors, which presents a significant challenge, particularly in the context of darknets. The inherent non-interactivity of darknets limits forensic capabilities and the depth of threat interpretation, as the passive nature of these systems restricts observable behavioral data, thereby hindering a comprehensive analysis of attacker intent.

In contrast, honeypot systems are actively deployed to attract and engage attackers, offering a higher degree of interactivity. By analyzing attacker behavior within a controlled environment, inferences regarding their motives can be drawn. For example, attempts to exploit known vulnerabilities suggest a desire for unauthorized system access, while data exfiltration attempts indicate information theft as a primary objective. Furthermore, the tools and techniques employed by attackers provide valuable insights into their overarching goals.

While both darknets and honeypots contribute to threat intelligence gathering, their distinct characteristics influence the nature and volume of collected data. Factors such as system configuration, data collection methodologies, and levels of interactivity all impact the depth of threat analysis [98]. Honeypots, although capable of providing detailed information on attacker tactics, techniques, and procedures (TTPs), face limitations due to high operational costs, which restrict deployment scale. Consequently, the number of

observed attackers and the diversity of attacker behaviors remain comparatively small. Darknets, on the other hand, offer a more scalable and cost-effective alternative through their passive, distributed sensor networks, but they are fundamentally limited by their non-interactive nature. This restricts the range of observable attacker behaviors and the amount of exploitable payload data available for detailed motive inference.

The reconnaissance phase of many attacks, characterized by network scanning, often precedes the actual exploitation phase. The specific scanning strategies employed reveal valuable information about attacker objectives. For instance, the selection of targeted ports for scanning indicates the types of systems and vulnerabilities sought by attackers. This selective targeting, combined with the use of evasion and persistence techniques, provides insights into the sophistication of the attacker and their overall goals. Therefore, the analysis of scanning traffic presents a rich source of information regarding threat actor motivation.

Consider a scenario in which a single threat actor is detected simultaneously by both a darknet and a honeypot deployed within the same geographical region. The honeypot, designed for exploitability, is likely to experience targeted attacks focused on commonly known vulnerabilities, resulting in a relatively limited scan range. Conversely, the darknet, characterized by its non-responsive nature and closed ports, induces a broader scanning effort by the attacker, who will likely utilize a more extensive exploit kit. The vast sensor network of the darknet and its continuous monitoring capabilities generate a more comprehensive scanning profile, while the honeypot provides detailed interactive logs that enable in-depth analysis of the threat actor's actions and motivations.

The correlation between these distinct datasets, obtained from both the darknet and honeypot, pertaining to the same threat actor, presents a powerful opportunity for enhanced threat intelligence. This correlation allows for mapping between the coarse-grained observations of the darknet and the fine-grained details provided by the honeypot, thereby linking darknet activity to underlying threat motives. This is particularly significant given the vast scale of darknet observation; large darknets like ORION are capable of observing millions of threat actors daily, significantly exceeding the capacity of honeypots, which typically observe only a few thousand. This disparity underscores the potential for leveraging darknet data to enhance early threat detection.

This dissertation investigates the comparative efficacy of network telescopes and honeypots in early threat detection, utilizing data from the ORION Network Telescope [1] and GreyNoise (GN) [99] - the honeypot used for this research. The analysis is predicated on two key observations.

First, a significant discrepancy exists in the volume of unique IP addresses detected by these two distinct threat intelligence sources. Figure 5.1 presents a comparative analysis spanning a 25-day period, revealing that the ORION darknet identified approximately 3.5 times more unique IP addresses (a difference of roughly 5,000,000) than GN. This disparity highlights the potential for increased coverage afforded by the larger observational aperture of the network telescope.

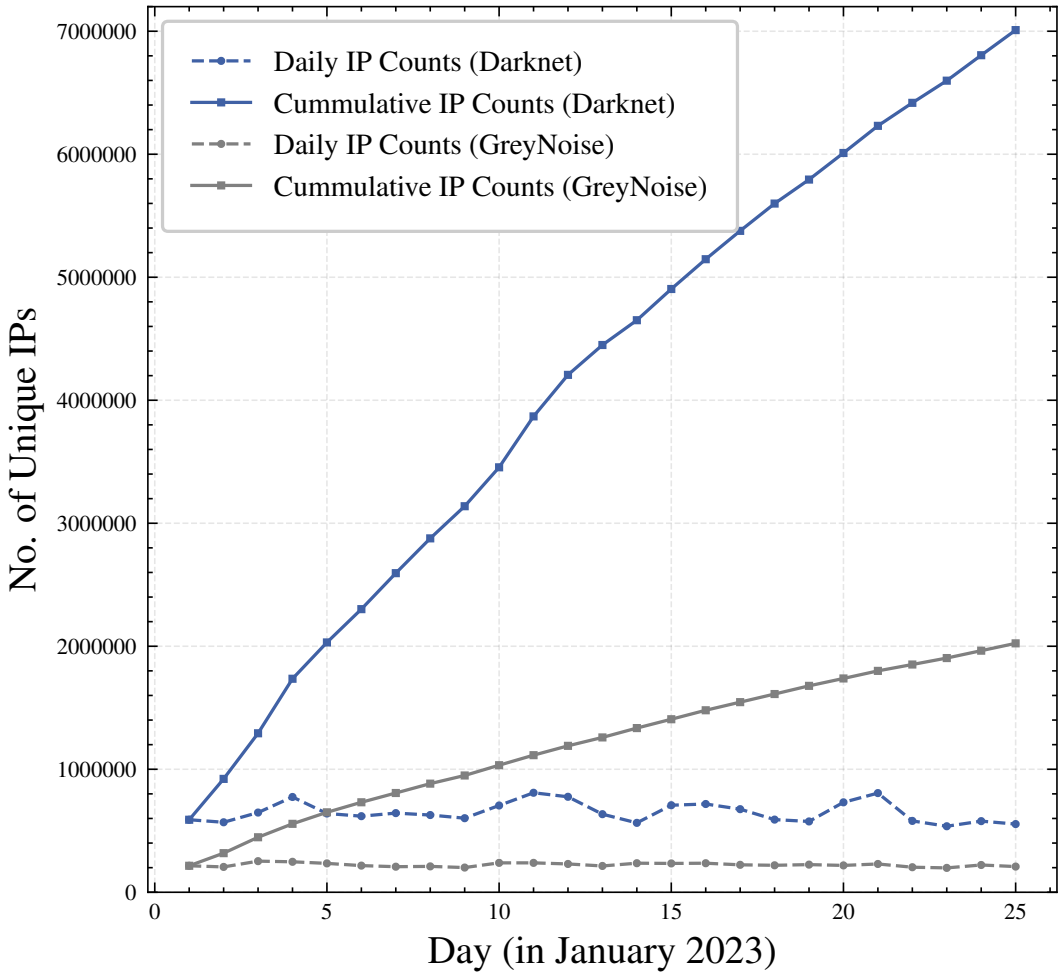


Figure 5.1: ORION darknet consistently records five times more observed IPs than GreyNoise, both daily and monthly.

Second, it is hypothesized that the increased observational scale of network telescopes translates to a reduction in threat detection latency. To test this hypothesis, the focus was placed on "common fresh IPs," defined as IP addresses that were observed for the first time within a given month by both the ORION Network Telescope and GN. For each common fresh IP, the temporal difference between its initial detection in ORION

and its initial detection in GN was calculated. This temporal difference was categorized as either "Darknet Lead Time" (when ORION detected the IP first) or "GreyNoise Lead Time" (when GN detected the IP first).

Figure 5.2 illustrates the daily average of these lead times. The results revealed a mean darknet lead time of approximately 12 hours, indicating that, on average, the network telescope detected common fresh IPs half a day earlier than GreyNoise. In contrast, the average GreyNoise lead time ranged from 2 to 5 hours. In contrast, the average GreyNoise lead time ranged from 2 to 5 hours, demonstrating a notably shorter detection window. This significant difference underscores the potential advantages of incorporating network telescope data into threat detection models.

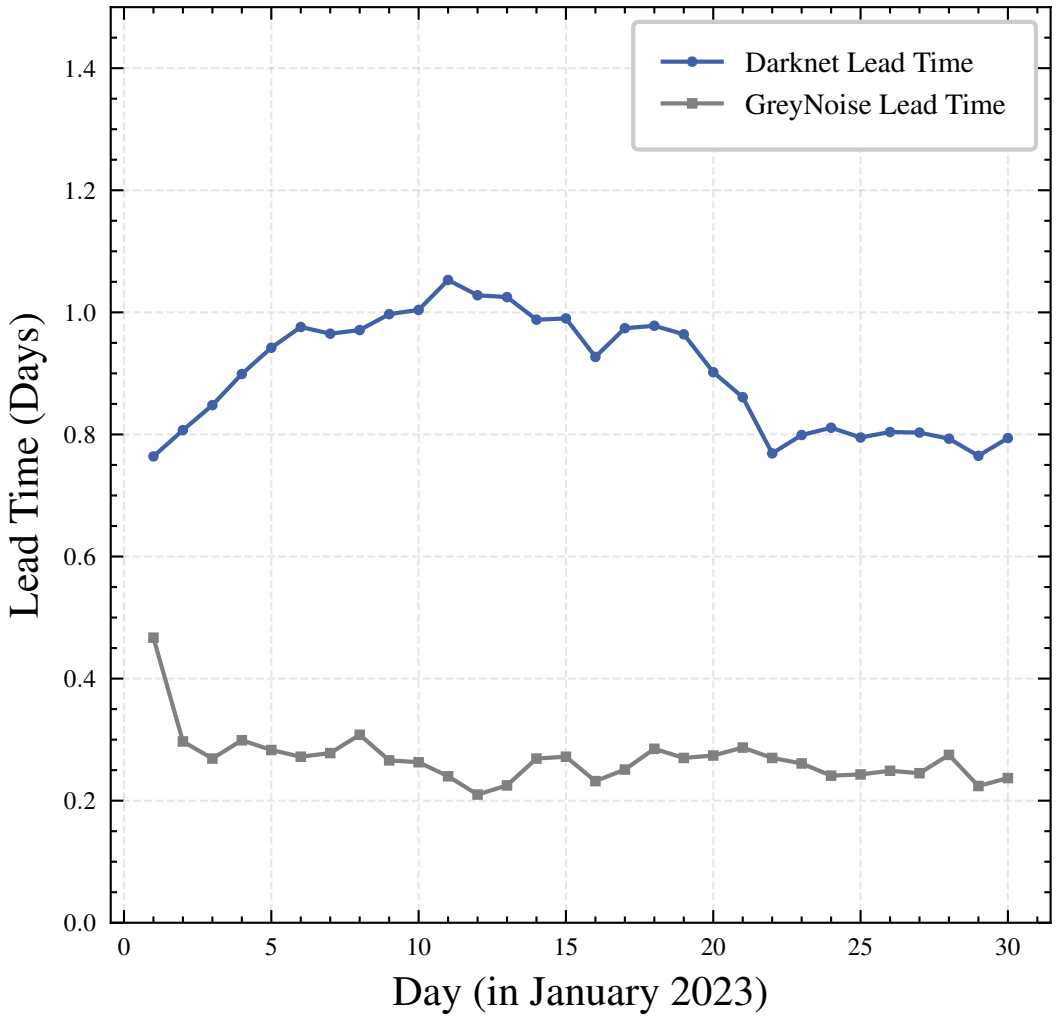


Figure 5.2: On average, ORION darknet detects common fresh IPs approximately 12 hours earlier than GreyNoise.

The implications of these findings indicate that the integration of data from network telescopes has the potential to significantly enhance the speed and efficiency of threat detection processes. By leveraging the extensive observational capabilities of darknets, security systems can identify emerging threats more rapidly, thereby facilitating timely intervention and improving overall cybersecurity resilience.

These findings suggest that a predictive model utilizing network telescope scanning behavior data can substantially enhance honeypot-based threat labeling and improve detection times for specific malicious actors. Future research will focus on the development and evaluation of such a model, incorporating advanced machine learning techniques to analyze the unique characteristics of network telescope data. This will aim to improve the accuracy and timeliness of threat detection and attribution.

Specific attention will be given to feature engineering techniques tailored to the high-volume, high-velocity nature of network telescope data, along with an investigation into model robustness and generalizability across diverse threat landscapes. Furthermore, the potential for integrating this model with existing Security Information and Event Management (SIEM) systems will be assessed, facilitating its seamless incorporation into operational security workflows. This integrated approach is expected to strengthen threat detection capabilities and enhance overall security posture.

5.2 Honeypot

This research is underpinned by two primary data sources. The darknet data is supplemented by threat intelligence derived from a distributed honeypot network. This network consists of strategically positioned sensors located across diverse geographical areas and cloud providers. The deployment is designed to attract malicious actors, thus enabling the capture of detailed information regarding their attack methodologies. This dataset encompasses a broad spectrum of attacker characteristics, including targeted vulnerabilities, deployed malware (such as worms), scanning tools, authentication attempts, and metadata related to crawlers, programming libraries, and search engines.

Signature-based detection is employed by the honeypot to identify patterns within the heterogeneous malicious traffic data. This process results in the generation of a comprehensive set of threat labels that characterize attacker behavior, traffic profiles, penetration and exploitation techniques, and overall attack intent. While the exact label generation process remains proprietary, the resulting threat labels provide valuable insights into attacker behavior. These labels facilitate the efficient assessment and

prioritization of risks by security analysts, enabling timely and effective responses.

The honeypot utilized for this research is operated by GreyNoise Intelligence, Inc. Specifically, the threat labels derived from the honeypot's infrastructure are integrated with scanning profiles collected by the ORION network telescope. This integration yields a labeled dataset suitable for the training of predictive models that map observed scanning patterns to threat labels. Example threat labels across five illustrative categories are presented in Table 5.1. The construction of this dataset is crucial for the subsequent machine learning phases described in Section 5.4. The integration strategy explicitly addresses the challenge of combining disparate data sources with varying levels of granularity and fidelity, a critical aspect contributing to the methodological novelty of this work.

Table 5.1: Exemplar threat labels from different categories.

Category	**Examples**
Scan	Arucer Crawler, Printer Crawler, Docker Scanner, IMAP Crawler
Exploit	GPON CVE-2018-10561 Router Worm, Realtek MiniiGD UPNP Worm CVE-2014-8361, EIR D1000 Router Worm, D-Link UPNP OS Command Injection, CCTV-DVR RCE, Vacron NVR RCE
Tools	Zmap, Nmap, Python Requests Client, Go HTTP Client
Malware	Mirai, Linksys Eseries The Moon Worm, Looks Like RDP Worm
Brute-Force	Telnet Bruteforcer, SSH Bruteforcer, Tomcat Manager Brute Force Attempt

5.3 Problem Formulation

Let D_N be a darknet's dataset of daily scanning profiles $< IP_D, f_1, f_2, ...f_n >$, where IP_D denotes source IP address of each scanner observed in the darknet, and f_i denotes i^{th} feature of the scanner. Let H_P be the dataset of threat labels from a honeypot for the same day $< IP_H, l_1, l_2, ..., l_m >$, where IP_H denotes source IP address of each threat actor observed by the honeypot, and l_j denotes the j^{th} threat label. The integration of the two datasets, based on common source IP, results in an integrated dataset: $< IP_D, f_1, f_2, ..., f_n, IP_H, l_1, l_2, ..., l_m >$, where $IP_D = IP_H$.

Each scanner identified within the ORION darknet is characterized by a high-dimensional feature vector $f \in \mathbb{R}^P$. An autoencoder is employed to facilitate the

encoding of this high-dimensional feature space into a lower-dimensional representation, transforming f into an embedded vector $x \in \mathbb{R}^Q$, where $Q \ll P$, while preserving essential information from the original vector. The efficacy of the resulting vector representation in capturing critical information from the original data space has been demonstrated in both clustering and temporal change detection tasks, as outlined in [42].

This embedded feature vector from the darknet for the scanners, combined with the corresponding labels derived from Honeypot data, constitutes an autoencoded Multi-Label Dataset (MLD), $M = \{(x_i, Y_i)|i = 1, \ldots, n\}$, where Y_i denotes the threat labels associated with the i^{th} scanner, and $n = |M|$ represents the total number of multi-label instances, each corresponding to an individual scanner.

The challenge of multi-label classification is to identify a function F that effectively maps the embedded feature vectors x_i to the corresponding threat label set Y_i as shown by:

$$F(x_i) = \hat{Y}_i \tag{5.1}$$

where $\hat{Y}_i \subseteq L$ is the set of predicted labels.

Upon the completion of training for the classifier F, it can be employed to infer threat labels for the set of IP addresses observed in the darknet that have not yet been encountered in the honeypot. This application of the model facilitates the amplification of threat intelligence and enhances the timeliness of detecting dubious activities. Given that the label set L may evolve over time due to the emergence of new vulnerabilities, malware, and other security threats, the model F is subject to periodic retraining. This retraining process is essential for ensuring the model's continued relevance and accuracy in the dynamic landscape of cybersecurity, as will be elaborated upon in Section 5.7.

5.4 Construction of the Integrated Dataset

The central element of the proposed framework involves the establishment of a mapping from the scanning features logged by ORION darknet to the comprehensive labels containing detailed threat characteristics observed by GreyNoise. This mapping is trained using a set of scanning IPs that are recorded by both data sources. It is hypothesized that an association exists between the data recorded for these shared IPs, under the assumption that the IP represents the same device and behavior across different sensors. However, the dynamic nature of IP assignments and the evolving behaviors of malicious actors present specific challenges to this assumption. Therefore, a short-time window of $\Delta t = 1\,\text{day}$ is adopted, during which it can be assumed with higher confidence that an IP

observed in both ORION and GreyNoise refers to the same scanning device functioning with the same threat characteristics. A day-length window has also been commonly utilized as a plausible time period for IP address-device stability in other works [100].

To establish a predictive model that effectively maps network scanning features from darknet to threat labels from honeypot, it is essential to construct a unified dataset that integrates these disparate data sources. This integration process, guided by the methodology outlined in Section 5.2, initiates with the identification of source IP addresses that are concurrently observed by both data sources on the same calendar days. Following this, a join operation is performed to merge the scanning profiles with the corresponding threat labels associated with each shared IP address.

In instances where a single source IP may be linked to multiple threat labels, the resulting dataset is structured as a multi-label dataset, allowing each scanning instance to be concurrently assigned one or more threat labels. The inherent multi-label nature of the dataset necessitates a nuanced approach to both data analysis and model development. Several characteristics unique to multi-label data must be thoroughly analyzed prior to making any inferences. Section 5.4.1 briefly considers some of the key characteristics of this multi-label dataset.

5.4.1 Characteristics of Multi-Label Dataset

Effective predictive modeling on this multi-label dataset (MLD) necessitates a comprehensive understanding of label interdependencies and correlations. The relationships between labels, including patterns of label co-occurrence (illustrated in Figure 5.3), alongside the prevalence of class imbalance across various threat labels, present significant challenges to the training of models. Careful consideration and implementation of mitigation strategies are required to address these complexities. A comprehensive overview of techniques for addressing class imbalance and label correlated in multi-label datasets has been considered in this study in the context of multi-label learning.

5.5 Multi-Label Learning

Multi-label learning (MLL) is recognized as a specialized area within machine learning that addresses instances associated with multiple labels simultaneously, in contrast to traditional single-label learning, where each instance is linked to only one label. The objective in multi-label learning is to accurately predict the correct set of labels for a given

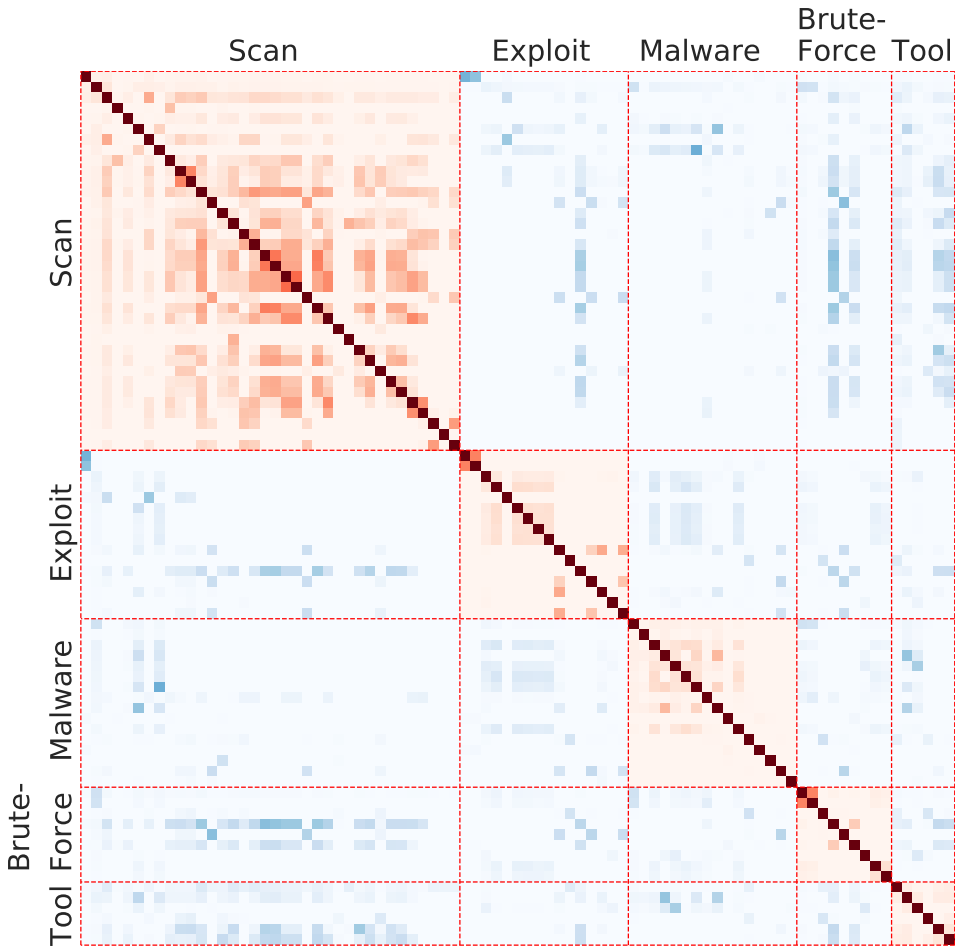


Figure 5.3: Concurrence among the labels. Each row/column represents a label. Darker (more saturated) colors indicate high degree of concurrence.

instance. Multi-label classification (MLC) represents a specific subtype of multi-label learning, characterized by predefined and finite labels, with the goal of assigning the appropriate subset of labels to each instance in the dataset.

Let $\mathcal{X} = \mathbb{R}^d$ be the d-dimensional instance space and $\mathcal{Y} = \{y_1, y_2, \ldots, y_q\}$ be the label space with q possible class labels. The objective of multi-label learning is to learn a function $f : \mathcal{X} \to 2^{\mathcal{Y}}$ from the multi-label training set $D = \{(\mathbf{x}_i, \mathbf{Y}_i)|1 \leq i \leq m\}$, where m is the total number of instances, $\mathbf{x}_i \in \mathcal{X}$ is a d-dimensional feature vector $(x_{i1}, x_{i2}, \ldots, x_{id})$, and $\mathbf{Y}_i \subseteq \mathcal{Y}$ is the set of labels associated with $\mathbf{x}_i$. In most cases, f is a real-valued function $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, where $f(\mathbf{x}, y)$ represents the confidence of $y \in \mathcal{Y}$ being the proper label of $\mathbf{x}$. Specifically, $f(\cdot, \cdot)$ should output larger values for the relevant label $y \in \mathcal{Y}$ and smaller values for the irrelevant label $y' \notin \mathcal{Y}$, i.e. $f(\mathbf{x}, y) > f(\mathbf{x}, y')$ for any given multi-label example $(\mathbf{x}, y)$. A thresholding function $t : X \to \mathbb{R}$ is applied to

$f(\cdot,\cdot)$ to create a multi-label classifier $h(\cdot)$, where $h(\mathbf{x}) = \{y|f(\mathbf{x},y) > t(\mathbf{x}), y \in \mathcal{Y}\}$. The threshold function divides the label space into relevant and irrelevant label sets. Given any unseen instance $\mathbf{x} \in \mathcal{X}$, $h(\cdot)$ predicts $h(\mathbf{x}) \subseteq \mathcal{Y}$ as the set of proper labels for $\mathbf{x}$.

The form of the functions $f(\cdot,\cdot)$ and $h(\cdot)$ utilized depends on the particular algorithms employed, and given that multi-label datasets (MLDs) are prevalent in diverse fields, such as medical diagnosis and image/video annotation, there is a vast literature on algorithms for multi-label learning and multi-label classification [101, 102].

Multi-label learning presents several challenges. One significant difficulty arises from the exponential growth of label combinations as the number of labels increases, resulting in label sparsity. This sparsity complicates the learning of accurate models without the risk of overfitting, particularly when most instances are associated with only a limited number of labels. Additionally, the presence of correlated labels is a common characteristic of multi-label datasets (MLDs), where the occurrence or absence of one label may be associated with the occurrence or absence of another label. This correlation can hinder the independent prediction of labels.

Another critical issue pertains to imbalanced label distribution, where certain labels appear significantly more frequently than others, leading to biased models that exhibit poor performance on minority labels. Furthermore, the presence of missing or noisy labels can pose additional challenges, as some instances may lack specific labels or contain ambiguous or erroneous labels, complicating accurate prediction. Failure to adequately address these challenges can result in inaccurate predictions, diminished model performance, and increased computational costs. To ensure that algorithms are accurate, efficient, and effective in solving real-world problems, most modern algorithms are inherently designed to handle these complexities.

5.5.1 Label Correlation

Multi-label learning addresses the challenge of assigning multiple labels to a single instance, a scenario frequently encountered in various real-world applications. A critical aspect of MLL lies in the inherent correlation between labels; the presence or absence of one label significantly influences the probabilities associated with other labels. This interdependence, quantified as label correlation, may be classified as either positive (where the presence of one label increases the likelihood of another) or negative (where the presence of one label decreases the likelihood of another). Neglecting to account for label correlation can lead to suboptimal model performance, as the independence assumption that underpins many standard MLL algorithms is violated. Consequently,

predictions may become inconsistent with the true underlying dependencies, resulting in diminished accuracy and predictive power.

The impact of label correlation on predictive accuracy is multifaceted. For example, in a medical diagnosis scenario, the presence of a symptom (label) may strongly indicate the likelihood of a particular disease (another label). Ignoring this positive correlation could result in missed diagnoses or inaccurate treatment plans. Conversely, in image classification, the presence of one object (label) may preclude the presence of another mutually exclusive object (negative correlation), necessitating sophisticated modeling to accurately reflect such dependencies. Thus, effectively addressing label correlation is deemed paramount for the construction of robust and accurate multi-label learning models.

Various strategies have been developed to incorporate label correlation into MLL algorithms. These approaches can be categorized hierarchically based on the order of label interactions considered: first-order, second-order, and high-order strategies. First-order strategies, such as those utilizing naive Bayes classifiers, implicitly assume label independence and consequently fail to explicitly model label correlations. In contrast, second-order strategies explicitly model pairwise label dependencies. These methods often leverage techniques such as graphical models, including Bayesian networks or Markov random fields, to represent the conditional dependencies between label pairs, conditioned on the input features. This approach allows for a more nuanced understanding of label relationships and subsequently enhances predictive capabilities.

High-order strategies extend this capability to model interactions among multiple labels simultaneously. These approaches are characterized by increased computational complexity but can capture intricate dependencies beyond pairwise relationships. Methods such as tensor factorization or higher-order Markov random fields can be employed to manage such high-order interactions. The choice of strategy is influenced by the complexity of label relationships and the computational resources available. For datasets with intricate label dependencies, high-order methods may yield significant performance improvements. However, for datasets primarily exhibiting pairwise dependencies, second-order methods may provide an optimal balance between performance and efficiency.

5.5.2 Imbalanced Label distribution

Multi-label learning presents unique challenges, particularly the widespread issue of imbalanced label distributions. Unlike uniformly distributed datasets, multi-label datasets often exhibit significant disparities in label frequencies, with some labels having numerous

positive instances while others are sparsely represented. This inherent characteristic, prevalent across many MLL domains, can substantially compromise the performance of standard classification algorithms. Specifically, models trained on such data are prone to bias towards majority class labels, leading to suboptimal performance — and potentially unacceptable error rates — on under-represented minority labels. This bias results in a diminished capacity to effectively learn the intricate relationships associated with minority labels, producing a model that is unreliable for predicting these less frequent yet potentially critical labels.

Several quantitative metrics have been proposed to characterize this label imbalance. One such metric is the Imbalance Ratio per Label (IRLbl), which offers a label-specific assessment of imbalance. The IRLbl for a given label is computed as the ratio of the frequency of the most frequent label to the frequency of that label. A higher IRLbl value indicates a greater degree of imbalance for that specific label. This granular analysis facilitates the identification of individual labels exhibiting extreme imbalance, allowing for the implementation of targeted mitigation strategies. Furthermore, aggregate metrics derived from IRLbl, such as the Mean Imbalance Ratio (MeanIR), Maximum Imbalance Ratio (MaxIR), and Coefficient of Variation of IRLbl (CIR), provide a comprehensive overview of the overall dataset imbalance

Addressing label imbalance is a well-established problem in single-label classification, typically approached through resampling techniques, such as oversampling minority classes or undersampling majority classes, or through the application of inherently robust algorithms, such as decision trees or ensemble methods. However, the direct application of these single-label strategies to the multi-label setting often proves problematic. While resampling methods can be effective in single-label classification, they may inadvertently introduce spurious label dependencies into the data, distorting the original relationships between labels. This issue is particularly relevant in multi-label learning, where the co-occurrence of labels is a key characteristic. In highly imbalanced datasets, majority labels frequently co-occur with minority labels, a phenomenon that resampling techniques could significantly alter.

This co-occurrence of labels, referred to as label concurrence, necessitates careful consideration when mitigating imbalance. The high probability of a majority label appearing alongside a minority label in imbalanced MLL datasets requires a nuanced approach. Simple resampling methods may disrupt this inherent concurrence, potentially resulting in a model that fails to capture crucial label relationships. To quantify label concurrence, the SCUMBLE (SCore for Unbalanced Multi-Label dAtasEts) score is

introduced, a metric explicitly designed to capture the level of label co-occurrence in unbalanced multi-label datasets. The formal definition of the SCUMBLE score is provided below:

$$\text{SCUMBLE}(D) = \frac{1}{n}\sum_{i=1}^{n}\left[1 - \frac{1}{\overline{IRLbl_i}}\left(\prod_{\lambda\in\mathcal{Y}} IRLbl_{i\lambda}\right)^{\frac{1}{|\mathcal{Y}|}}\right] \tag{5.2}$$

where n is the number of samples in the dataset, $\mathcal{Y}$ represents the set of all labels in the dataset, and $\overline{IRLbl_i}$ is the average imbalance level of the labels that appear in the i-th sample. $IRLbl_{i\lambda}$ is equal to $IRLbl(\lambda)$ if $\lambda \in Y_i$; otherwise $IRLbl_{i\lambda} = 0$. The imbalance ratio $IRLbl(\lambda)$ for a label λ is defined as:

$$\begin{aligned} IRLbl(\lambda) &= \frac{\max(\sum_{i=1}^{n} h(\lambda', Y_i))}{\sum_{i=1}^{n}(h(\lambda, Y_i))} \\ h(\lambda, Y_i) &= \begin{cases} 1, & \lambda \in Y_i \\ 0, & \lambda \notin Y_i \end{cases} \end{aligned} \tag{5.3}$$

The SCUMBLE score measures the imbalance variance among the labels present in each sample. It quantifies the degree of concurrency among imbalanced labels, with higher scores indicating higher concurrency and vice versa.

The presence of significant class disparities, wherein certain labels occur far more frequently than others, can severely compromise the performance of standard learning algorithms. This issue is particularly problematic when high label co-occurrence exists, a phenomenon characterized by the frequent appearance of minority labels alongside majority labels. The strategy of oversampling minority classes to address this imbalance, commonly employed in binary classification, may inadvertently exacerbate the problem within the multi-label context. Specifically, oversampling a minority label may increase the frequency of majority labels that frequently co-occur with it, thereby further skewing the overall label distribution [103]. Conversely, undersampling majority labels poses the risk of removing instances associated with already scarce minority labels, potentially hindering the learning process for those underrepresented classes. This effect is amplified when the co-occurrence of minority and majority labels is substantial.

The limitations inherent in simple resampling techniques necessitate the exploration of more sophisticated approaches to handle imbalanced multi-label datasets. Random oversampling and undersampling, while effective in binary classification scenarios with independent classes, are often found to be insufficient and may even worsen the problem in multi-label settings characterized by high label co-occurrence. The potential for increased

noise in the resampled data, which is introduced by artificially generating samples that may not accurately reflect the underlying data distribution, further undermines these naive approaches. Consequently, advanced techniques are required to effectively mitigate the detrimental effects of label imbalance.

Methods such as MLSMOTE (Multi-label Synthetic Minority Over-sampling Technique) [104] have been developed to address these shortcomings by generating synthetic minority class instances that intelligently consider label co-occurrence patterns. Unlike random oversampling, MLSMOTE aims to create synthetic samples that better reflect the underlying relationships between labels, thereby reducing the risk of introducing spurious correlations and noise into the dataset. This approach circumvents the pitfalls associated with randomly generating samples, which could exacerbate the existing imbalance and lead to overfitting.

Beyond resampling, alternative strategies for addressing imbalanced multi-label datasets include cost-sensitive learning, which assigns varying misclassification costs to different labels based on their prevalence. By imposing heavier penalties for misclassifications of minority labels, cost-sensitive learning seeks to enhance the model's ability to identify these underrepresented classes. Additionally, adaptive threshold adjustment techniques can be employed to fine-tune the decision boundaries of the classifier, enabling more nuanced classification of instances associated with minority labels. Algorithm adaptation involves modifying the learning algorithm itself to better accommodate imbalanced data, while ensemble methods combine predictions from multiple classifiers trained on different subsets of the data or under varying learning conditions to improve overall robustness and reduce bias.

This chapter of the dissertation investiates multi-label learning algorithms in relation to the problem of cybersecurity threat inference. Multi-label learning is particularly relevant due to the inherently multifaceted nature of cyber threats, wherein a single incident may manifest with multiple associated labels that represent diverse attack vectors, targets, and consequences. This section systematically explores the methodological landscape of MLL, with a focus on the selection and optimization of algorithms tailored for this specific application.

5.5.3 Learning Algorithm Selection and Categorization

Two primary paradigms dominate the MLL literature [105]: problem transformation and algorithm adaptation. Problem transformation methods recast the MLL problem into a more conventional machine learning framework, such as binary classification, label

ranking, or multi-class classification. Examples include Binary Relevance (BR) [106], which decomposes the problem into independent binary classifications for each label, and Classifier Chains (CC), which sequentially chains classifiers to model label dependencies. Alternatively, algorithm adaptation methods modify existing single-label learning algorithms to directly handle multiple labels. Examples include adaptations of k-nearest neighbors (ML-kNN), decision trees (ML-DT), support vector machines (e.g., Rank-SVM), and information-theoretic approaches (e.g., CML).

Problem transformation methods offer advantages in terms of simplicity and interpretability. BR, for instance, is straightforward to implement but suffers from the strong assumption of label independence. CC mitigates this limitation by incorporating label correlations through a chained classifier structure, though at the cost of increased computational complexity, particularly with a large number of labels. Algorithm adaptation, conversely, can offer computational efficiency, especially when dealing with imbalanced label distributions, by directly addressing the multi-label nature of the data within the algorithm's core structure. The choice between these paradigms necessitates a careful consideration of the dataset characteristics (size, label sparsity, imbalance) and the trade-off between computational cost, model interpretability, and predictive performance.

5.5.3.1 Ensemble Methods and Advanced Techniques

Ensemble methods enhance the performance of MLL models by combining the predictions of many base classifiers. These ensembles can effectively capture label correlations and leverage the strengths of diverse base learners (e.g., decision trees, support vector machines, k-nearest neighbors). Common ensemble construction techniques include "one-vs-all," where each base classifier is trained to predict a single label, and "error-correcting output codes," which employ binary codes to represent label combinations.

The increasing scale and complexity of modern datasets necessitate the adoption of even more advanced MLL techniques. Extreme Multi-Label Classification (XMC) [107] represents the domain of multi-label classification that addresses the challenges associated with large label spaces. Specialized approaches, such as sparse linear models, neural networks, and tree-based models, are required to tackle these challenges, often incorporating techniques like label embedding, hierarchical classification, and label pruning to effectively manage the expansive label space.

5.5.4 Model Selection and Hyperparameter Optimization

This research evaluates state-of-the-art multi-label classification algortihms from several classes: Classifier Chains [108], Random k-Label Sets (RAKEL) [109], Multi-Label Weighted Subspace Ensemble (MLWSE) [110], NapkinXC [111], and ProXML [112]. These algorithms are selected considering their inherent ability to address label correlation and imbalanced labels. Each algorithm necessitates careful hyperparameter tuning to achieve optimal performance.

All models were evaluated using 10-fold cross-validation with varying random seeds to ensure robustness. The superior model was selected based on its overall performance across multiple evaluation metrics. The trade-off between predictive performance, computational complexity (including training and inference time), and resource consumption was carefully considered.

5.5.4.1 Classifier Chains

Classifier chains employs a linking mechanism between binary base classifiers to model label dependencies. Its capability to handle label imbalance is noteworthy, as it does not presuppose an even distribution of labels. For CC, a 5-fold cross-validation on the training set was used to select the optimal base classifier from amongst Support Vector Classifier (SVC), Naive Bayes Classifier (NBC), and Logistic Regression Classifier (LRC). LRC emerged as the superior choice. The efficacy of classifier chains may be hindered if the ordering of base classifiers is sub-optimal, which can negatively impact performance. However, the ordering of base classifiers was deemed non-essential in this case based on preliminary experimentation.

5.5.4.2 RAndom k-labELsets (RAKEL)

The Label Powerset is a class of multi-label classification algorithms that treats each combination of labels as a distinct class, reducing the problem to predicting a single label from a finite set. This approach captures label correlation by considering all feasible combinations as separate classes, but performance challenges arise when there are insufficient instances for certain combinations. To address this limitation, the RAKEL algorithm [109] is employed, which partitions the label set into smaller, random subsets and constructs an ensemble of single-label classifiers, each trained on a specific label subset. This method effectively mitigates the issue of inadequate instances per label by focusing on a limited number of labels.

RAKEL requires the specification of the label subset size (k), the number of models, and the output threshold. A subset size of 3, shown to be effective in previous research, was adopted. The number of models was determined based on the total number of labels and the specified subset size. An output threshold of 0.5 was applied for the final prediction. Furthermore, through 5-fold cross-validation, ML-kNN was identified as the optimal base classifier for RAKEL, outperforming both LRC and SVC.

5.5.4.3 Multi-Label Stacked Ensemble (MLWSE)

In light of the efficacy of ensemble models, an innovative technique known as the Multi-Label Weighted Stacked Ensemble (MLWSE), as described in [110], has been integrated into this research. This approach not only facilitates the acquisition of weights for ensemble members but also leverages label correlations. In the implementation, a stacked ensemble comprising three base classifiers—Binary Relevance (BR), Classifier Chain (CC), and Label Powerset (LP)—was constructed, with the weights of the ensemble members computed by utilizing pairwise label correlations. The optimization algorithm introduced in [110] is employed to achieve the optimal amalgamation of the base classifiers and their respective weights within the ensemble.

5.5.4.4 Extreme Multi-Label Classification

With the emergence of novel malicious activities, the expansion of the GreyNoise label repository is anticipated, thereby posing a challenge for the aforementioned methods to accurately identify relevant labels. To address this issue, the performance of two state-of-the-art extreme multi-class (XMC) techniques, namely NapkinXC [111] and ProXML [112], is appraised in this study. NapkinXC is recognized as a rapid XMC methodology that employs probabilistic label trees. In parallel, ProXML is characterized as an optimization framework specifically designed to enhance tail label prediction in scenarios where the label count is substantial.

NapkinXC's key hyperparameter is the choice of solver for large-scale regularized classification. Experimentation with a library of linear solvers [111] led to the selection of liblinearSolver. ProXML utilized the best-performing model configuration described in Babbar and Schölkopf [112].

5.6 Evaluation

5.6.1 Evaluaton Metrics

5.6.1.1 Example-based Metrics

Example-based measures are first computed individually for each sample, then averaged to obtain the final value. If Y is the set of original labels and Z is the set of predicted labels, the following formulas are used to determine example-based metrics.

Precision measures the proportion of correctly predicted labels to the total number of predicted labels, averaged over all instances in the MLD.

$$Precision = \frac{1}{m}\sum_{i=1}^{m}\frac{|Y_i \cap Z_i|}{Z_i} \tag{5.4}$$

Recall is calculated as the ratio of correctly predicted labels to the total number of true labels, averaged over all samples in the MLD.

$$Recall = \frac{1}{m}\sum_{i=1}^{m}\frac{|Y_i \cap Z_i|}{Y_i} \tag{5.5}$$

F-measure, also called F1 score, combines recall and precision to provide a single weighted metric that assesses the amount of relevant labels that are predicted and the amount of predicted labels that are relevant. The most basic implementation of F-measure is the equally weighted harmonic mean of precision and recall as given by:

$$F\text{-}measure = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{5.6}$$

The higher the value of accuracy, precision, recall and F-measure, the better is the performance of the learning algorithm.

5.6.1.2 Label-based Metrics

Unlike example-based measures which are computed for each instance, label-based measures are calculated for each label. These measures first evaluate a known metric such as precision, recall, F-measure, etc. for each individual labels and these separately calculated metrics are then aggregated using averaging operations like macro averaging (measures are computed on individual labels first and then averaged over all labels) and micro-averaging (measures are calculated globally over all instances and all labels) [105].

If $FP\lambda$, $TP\lambda$, $FN\lambda$ and $TN\lambda$ denote the number of false positives, true positives, false negatives, and true negatives respectively and EM denote the known evaluation metric, then, the macro and micro averaged label-based measures can be computed as follows:

$$EM_{macro} = \frac{1}{|L|} \sum_{\lambda=1}^{|L|} EM(TP\lambda, FP\lambda, TN\lambda, FN\lambda) \tag{5.7}$$

$$EM_{micro} = EM(\sum_{\lambda=1}^{|L|} TP\lambda, \sum_{\lambda=1}^{|L|} FP\lambda, \sum_{\lambda=1}^{|L|} TN\lambda, \sum_{\lambda=1}^{|L|} FN\lambda) \tag{5.8}$$

5.6.2 Results

In this section, a comprehensive performance analysis of the selected multi-label classifiers is conducted, with evaluations performed across diverse label-based, example-based, and ranking-based metrics. Label-based metrics, encompassing precision, recall, and F1 scores, are initially computed for individual labels and subsequently aggregated over all labels using various averaging operations—micro, macro, and weighted.

After creating a multi-label dataset by combining feature profiles from ORION darknet with GreyNoise's labeled annotations for June 2023, an autoencoder is used to produce 50-dimensional embeddings for the input feature vectors. This process preserves essential information while capturing feature relationships, improving data representation and aiding subsequent analysis and classification tasks. The optimal autoencoder architectures, as indicated in [42], are replicated, asserting that a comprehensive and meaningful representation of scanning profile data can be encoded in a latent space of merely 50 dimensions without substantial information loss. The embeddings and labels are then formatted to align with the expected input format for each model; except for ProXML [112], which expects label indices. The labels are encoded using a multi-label binarizer before they are used.

Due to the imbalanced nature of the labels in this dataset, an oversampling technique (MLSMOTE algorithm [104]) is employed to generate 50,000 synthetic samples by oversampling the identified minority labels. This augmented data is utilized for the subsequent experiments. The evaluation results presented in Table 5.2 are averaged over these runs. The superior model among the mentioned approaches is selected based on its superior performance across the majority of the evaluation metrics described in Section 4.5. Additionally, a trade-off exists between performance and complexity in terms of training and inference time, as well as resource consumption, which must be considered when selecting a model for real-world threat inference applications.

Table 5.2: Comparison of Evaluation metrics across different classifiers.

Metrics	**Prec. Mac.**	**Rec. Mac.**	**F1 Mac.**	**Prec. Mic.**	**Rec. Mic.**	**F1 Mic.**	**Prec. Wtd.**	**Rec. Wtd.**	**F1 Wtd.**
Classifier chain [108]	0.82	0.81	0.81	**0.83**	0.85	**0.84**	0.8	**0.83**	0.81
MLWSE [110]	**0.85**	0.79	**0.82**	0.82	0.85	**0.84**	**0.82**	0.82	**0.82**
RAKEL [109]	0.78	**0.83**	0.80	0.78	**0.86**	0.81	0.77	**0.83**	0.80
NapkinXC [111]	0.75	0.78	0.76	0.78	0.76	0.77	0.65	0.76	0.70
ProXML [112]	0.73	0.79	0.76	0.78	0.83	0.80	0.74	0.73	0.73

MLWSE is identified as the most effective model, outperforming other models in 5 out of 9 evaluation metrics. Although the evaluation metrics employed consider each label to calculate the overall metric, it has been observed that this final metric can be misleading due to the imbalanced nature of the data. As illustrated in Figure 5.4, it is evident that while the model selected after experimentation demonstrates strong performance for common labels, a subset of threat labels is grossly mispredicted. An analysis of the areas where the model excels and where it fails is provided in Section 5.6.3.

5.6.3 Case Studies

5.6.3.1 Successful predictions

Given the distinctive port scanning patterns and traffic features exhibited by scanners and crawlers, our classifier demonstrates exceptional precision and recall in accurately predicting scanner-related labels, as illustrated in Figure 5.5. Decision trees [113], derived from the classifier's predictions, offer interpretable insights into its performance across different labels.

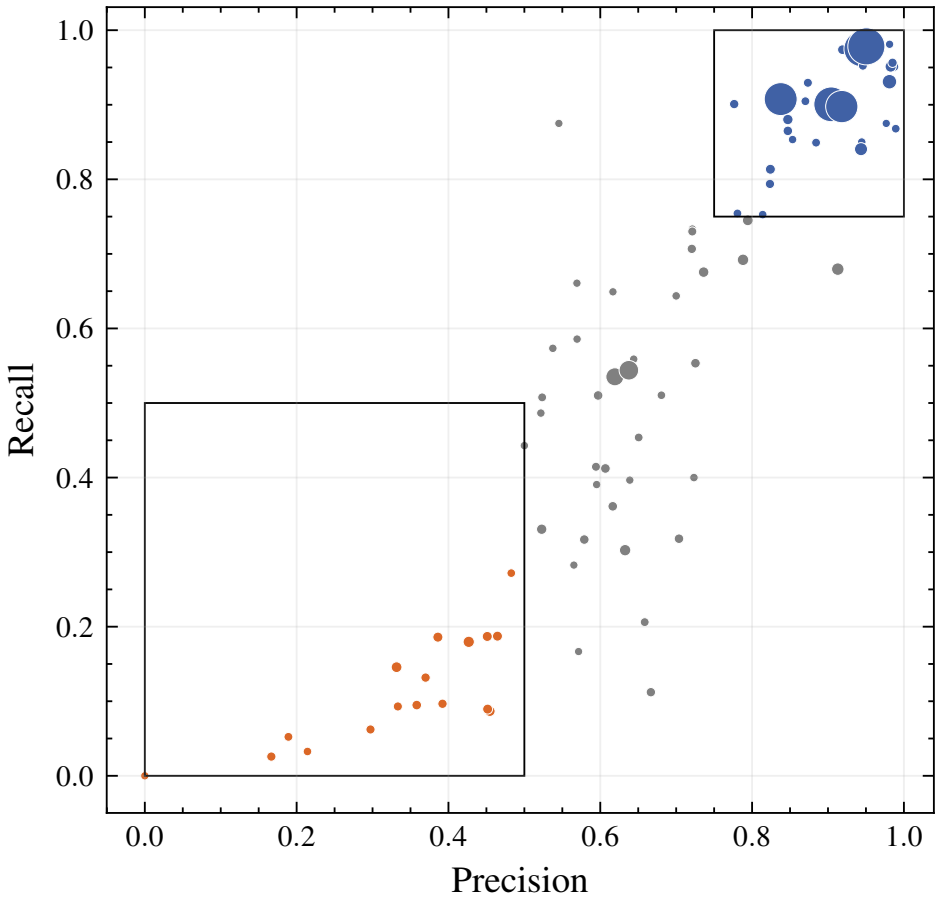


Figure 5.4: Bubble plot showing the prediction performance of the classifier chain on each label. The size of the bubbles is determined by the frequency of the label in dataset.

5.6.3.2 Difficulties in Model Prediction

The entities situated in the lower-left quadrant of the visual representation depicted in Figure 5.4 embody a circumstance characterized by the absence of suitable descriptors to elucidate a specific behavior. These designations encapsulate the strategies, methodologies, and processes undertaken by assailants in their attempts to infiltrate a system. Each label delineates a distinct vulnerability that the attacker endeavors to exploit within a particular device. These labels epitomize the threat intelligence amassed by a diminutive system, likely insufficient in comprehensively capturing the entire spectrum of behaviors. Consequently, as the attacker systematically targets diverse vulnerabilities, the honeypot may only observe a subset, thus attributing labels selectively. In this context, the model encounters challenges in assimilating knowledge from darknet features, which encounter the majority of the attacker's activities, with the honeypot's assigned labels.

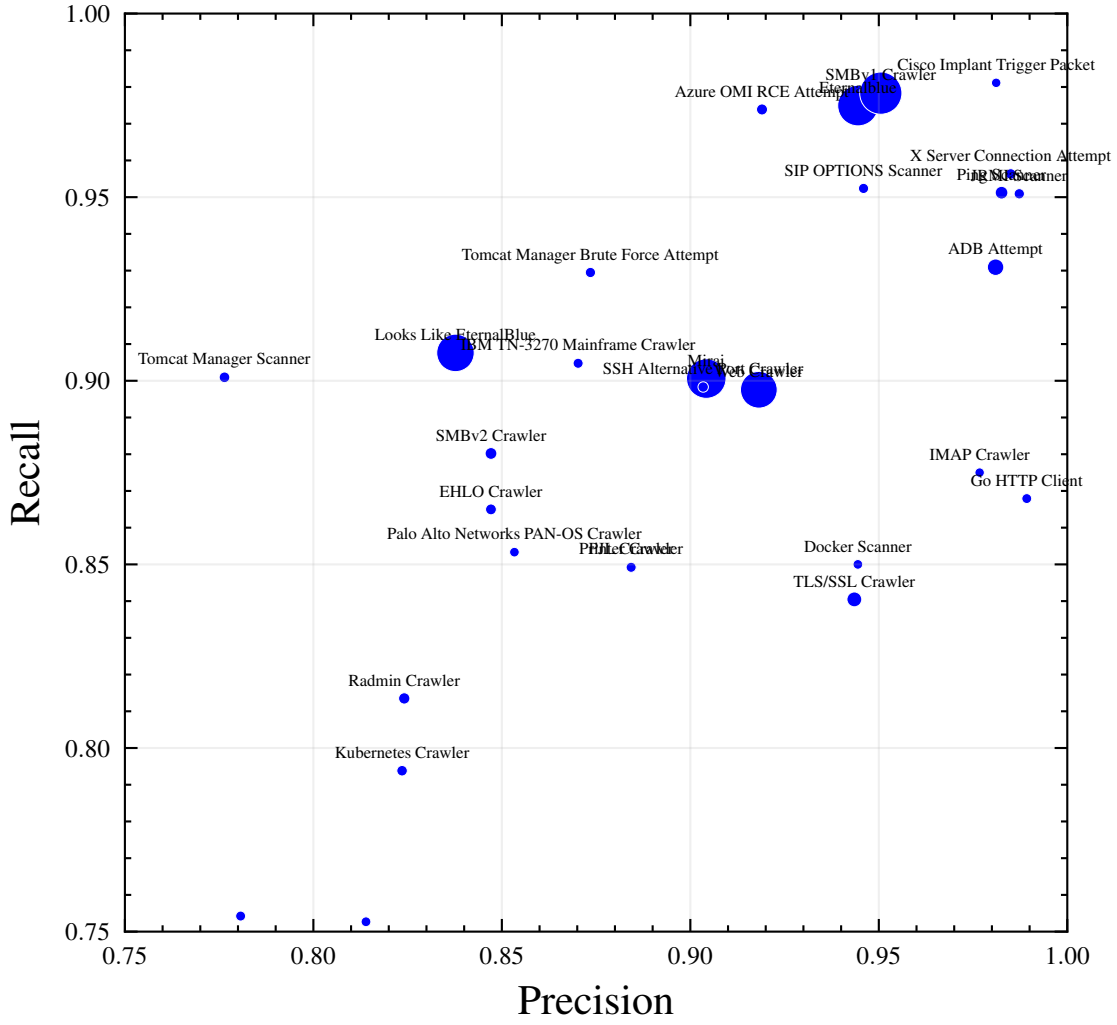


Figure 5.5: Bubble plot illustrating the predictive model's exceptional performance across a spectrum of labels, particularly excelling in identifying scanners, crawlers and connection attempts.

Furthermore, scrutiny of the payloads transmitted by these actors on the honeypot substantiates their affiliation with a botnet engaged in the dissemination of the Mozi malware. As our label repository lacked annotations for this malware during the study period, the model grapples with the inability to establish connections based on shared behaviors. This underscores the method's limitation in identifying zero-day attacks that exploit undisclosed or unknown vulnerabilities, yet to be disclosed by manufacturers.

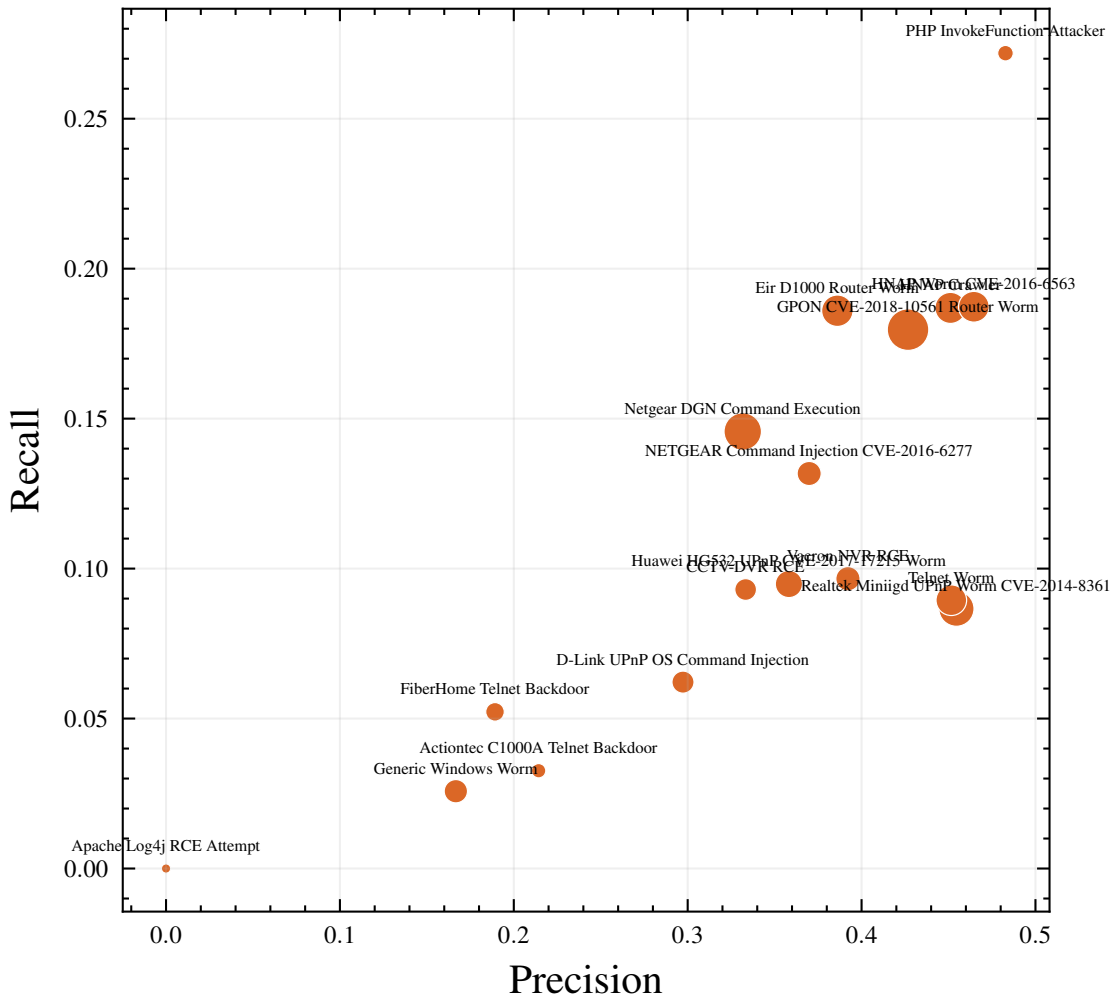


Figure 5.6: Bubble plot illustrating the predictive model's suboptimal performance on a subset of router-related exploits.

5.6.3.3 Router Exploits

Mozi, identified as a peer-to-peer (P2P) botnet, strategically capitalizes on unpatched vulnerabilities within IoT devices and exploits weak telnet passwords to compromise and infiltrate these devices [114]. Figure 5.6 illustrates instances where the classifier exhibits suboptimal prediction accuracy. Upon scrutinizing the IPs linked with this subset of labels and leveraging payload analysis, a discerning pattern emerged – all these instances were traced back to Mozi bots actively engaging in the dissemination and propagation of the Mozi source code. Noteworthy labels like "GPON CVE-2018-10561 Router Worm", "Realtek MiniiGD UPNP Worm CVE-2014-8361", "EIR D1000 Router Worm", and others correspond to explicit IoT vulnerabilities that these Mozi bots seek

to exploit, albeit their primary objective remains the infection of additional devices. The classifier faces a significant challenge due to the absence of a generic Mozi (*intent*) label, resulting in its limited ability to distinguish between these specific vulnerability exploits, which exhibit nearly identical traffic behavior and primarily differ in payload content.

5.7 Model Degradation and Retraining

The efficacy of machine learning models in network security, particularly within the context of network telescopes monitoring unused IP addresses, is significantly challenged by the inherent non-stationarity of network traffic data. This non-stationarity manifests as temporal variations in the statistical properties of the data, including its mean and variance. Furthermore, the dynamic nature of network behavior introduces both concept drift (changes in the underlying relationships between features and labels) and data drift (gradual shifts in the data distribution) [115]. These phenomena are amplified in the context of unused IP addresses, where the predominantly anomalous or malicious activity exhibits transient and rapidly evolving characteristics. The underlying causes of these drifts include the continuous adaptation of attack strategies, the shifting of attack targets, and the ever-changing network configurations. Therefore, robust model development necessitates addressing the multifaceted challenges posed by these temporal shifts.

The dynamic and non-stationary nature of network traffic necessitates a paradigm shift from static model training to a continuous retraining strategy to maintain model effectiveness. Models trained on historical data inevitably suffer from performance degradation over time, due to the divergence between the training data distribution and the evolving characteristics of real-time network traffic [116]. This temporal mismatch can lead to a significant increase in both false positives (incorrectly classifying benign traffic as malicious) and false negatives (failing to detect actual malicious activity). To mitigate these risks, a cyclical retraining regime, implemented on a daily, weekly, or monthly basis, depending on the observed rate of concept and data drift, is crucial. This continuous adaptation allows for the consistent refinement of detection capabilities and provides a robust defense against the evolving threat landscape, thereby maintaining high detection accuracy and minimizing the impact of both concept and data drift. The frequency of retraining should be determined empirically, based on the observed rate of change in the data distribution and the desired level of performance. Further research could investigate the optimal retraining frequency as a function of specific metrics related to concept and data drift detection.

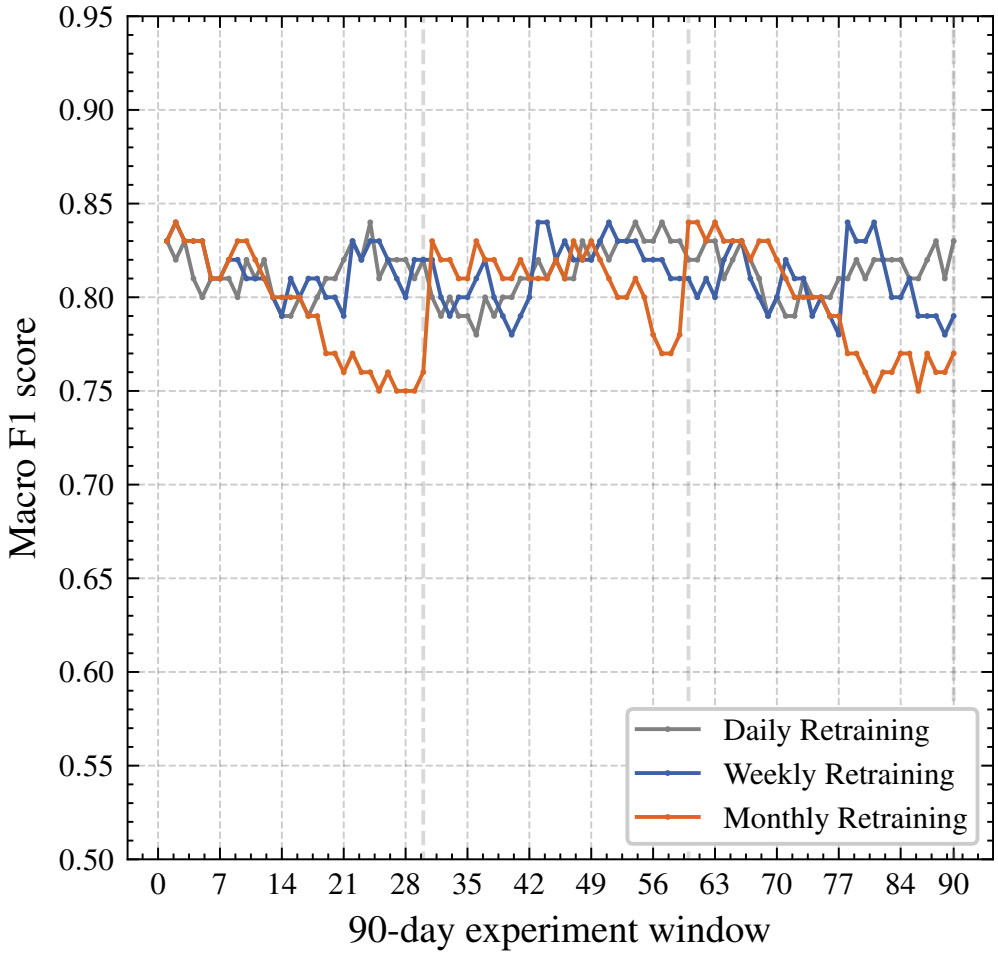


Figure 5.7: The similar performance observed among models retrained daily, weekly, and monthly during the initial two weeks suggests that bi-weekly training may be the optimal frequency for model training.

A comprehensive analysis was conducted on the retraining frequencies of the multi-label classifier using network traffic data collected over a 90-day period, with macro F1 scores serving as the performance metric. The findings, illustrated in Figure 5.7, indicate no significant performance differences between models retrained daily or weekly. Notably, the model retrained monthly exhibited performance comparable to that of the daily and weekly models during the first two weeks of the month, followed by a gradual decline thereafter. This suggests that a bi-weekly retraining frequency may be optimal, even though it was not included among the frequencies examined in this study, particularly when considering factors such as cost.

Importantly, the monthly retrained model occasionally outperformed the daily and weekly retrained models at the beginning of the month. These results reveal that different

models capture varying rates of data drift, with monthly retraining effectively addressing longer-term trends. Based on these observations, an ensemble approach [117] is proposed, which combines models retrained at different frequencies to leverage their complementary strengths and enhance overall performance in detecting and mitigating malicious activity within network traffic data.

An ensemble of machine learning models retrained at varying frequencies presents several advantages in the context of network traffic analysis. By utilizing models trained on different temporal scales, this approach enhances the ensemble's capacity to respond to diverse rates of data drift inherent in network traffic patterns. Specifically, the integration of models retrained at daily, weekly, and monthly intervals enables the ensemble to effectively capture both short-term fluctuations and long-term trends in the data, thereby improving overall detection performance.

To further optimize the ensemble's effectiveness, the implementation of a weighted approach based on the relative importance of each model's predictions is proposed [118]. In this framework, higher weights can be assigned to models that demonstrate greater efficacy in capturing prevailing data drift patterns. This dynamic weighting system can be adjusted over time according to the observed performance metrics of each model, allowing the ensemble to remain responsive to evolving data characteristics.

Additionally, the exploration of second-layer weight training for the ensemble is warranted, wherein the weights assigned to each model are learned directly from the data itself. This methodology holds the potential to enhance the ensemble's adaptability to complex and evolving data drift patterns, further improving its capacity for detecting and mitigating malicious activity within network traffic data. By implementing these strategies, the ensemble can achieve a more nuanced and responsive detection capability, thereby bolstering overall cybersecurity efforts.

5.8 Discussion

Adversaries are known for their proficiency in concealing their activities upon detecting certain systems, such as honeypots. When these systems serve as sources of enhanced threat intelligence, the labels assigned may be incomplete and insufficient to capture the full spectrum of malicious behaviors perpetrated by adversaries. In the context of the honeypot utilized in this investigation, specific incidents occurred where certain IPs were transiently detected, only to quickly depart, while more extensive probing efforts involving the same IPs were later observed in the darknet, despite its passive nature.

Two plausible explanations for this behavior can be posited: either the adversaries identified the honeypot's function and chose to evade it, or they successfully achieved their objectives during the initial interaction—eliciting the desired response from the sensor—and did not remain for additional actions. In the case of the darknet and the latter explanation, it is possible that subsequent scans followed the initial unsuccessful attempt. In both scenarios, relevant labels conveying attributes of the IPs remain undetected. Consequently, in datasets where such occurrences are prevalent, treating missing labels as inconsequential becomes justifiable; that is, the absence of a label may indicate either its irrelevance or its non-existence.

In datasets characterized by missing labels, employing a conventional approach to address the issue as a straightforward multi-label classification problem proves inadequate. Instead, this predicament necessitates an alternative strategy, akin to framing it as a weakly supervised multi-label learning problem with missing labels [119] or as positive unlabeled learning [120].

This study effectively demonstrates the practical applicability and utility of automated threat intelligence prediction for scanners, encompassing the identification of targeted vulnerabilities and the intended dissemination of malware. By leveraging the set of scanned ports in the darknet alongside other network-based darknet features, the predictive model, despite its inherent limitations, proves adept at generating meaningful threat intelligence labels for a diverse array of scanners. This capability bears significant implications for enhancing the cybersecurity posture of enterprises possessing unused IP spaces.

Enterprises can establish their own "Enterprise darknet" by monitoring network traffic directed at unused IP addresses or by analyzing firewall logs associated with networking attempts aimed at unused ports, similar to the methodology employed in Richter and Berger [85]. Subsequently, network-based profiles for observed scanners within this Enterprise darknet can be constructed. The mapping of network features to threat intelligence labels, as elucidated in this paper, can then be applied to these scanner profiles, revealing critical threat intelligence tailored specifically to scanners targeting the enterprise.

Moreover, the integration of this threat intelligence with the enterprise's firewall logs enables the intrusion detection team to swiftly identify potential endpoints within the enterprise network that have been contacted by scanners. This facilitates the initiation of timely and appropriate mitigation actions, thereby strengthening the cybersecurity defenses of the enterprise.

Learning using privileged information (LUPI) is a machine learning paradigm that leverages information available in the training data, but not available in the testing data. This additional information, often referred to as privileged information, can be used to enhance the learning process and improve the model's performance. LUPI machine learning models and theory were introduced by Vapnik and Vashist [121, 122].

Table 5.3: Comparison of Evaluation metrics across different classifiers trained on both available and privileged information.

	Prec. Mac.	**Rec. Mac.**	**F1 Mac.**	**Prec. Mic.**	**Rec. Mic.**	**F1 Mic.**	**Prec. Wtd.**	**Rec. Wtd.**	**F1 Wtd.**
Classifier chain [108]	0.91 (↑0.09)	0.90 (↑0.09)	0.90 (↑0.09)	**0.88** (↑0.05)	**0.89** (↑0.04)	**0.88** (↑0.03)	0.90 (↑0.06)	0.91 (↑0.08)	0.90 (↑0.09)
MLWSE [110]	**0.92** (↑0.07)	**0.93** (↑0.14)	**0.92** (↑0.10)	**0.88** (↑0.06)	0.86 (↑0.01)	0.87 (↑0.03)	**0.92** (↑0.12)	**0.92** (↑0.06)	**0.92** (↑0.09)
RAKEL [109]	0.88 (↑0.09)	0.91 (↑0.14)	0.89 (↑0.09)	**0.88** (↑0.07)	0.84 (↑0.05)	0.86 (0.00)	0.89 (↑0.05)	0.89 (↑0.06)	0.89 (↑0.16)
NapkinXC [111]	0.86 (↑0.11)	0.84 (↑0.06)	0.85 (↑0.09)	0.86 (↑0.08)	**0.89** (↑0.13)	0.87 (↑0.10)	0.84 (↑0.19)	0.81 (↑0.05)	0.82 (↑0.12)
ProXML [112]	0.85 (↑0.08)	0.86 (↑0.07)	0.85 (↑0.09)	0.86 (↑0.08)	**0.89** (↑0.06)	0.87 (↑0.13)	0.82 (↑0.08)	0.86 (↑0.13)	0.84 (↑0.11)

The integration of threat labels from Honeypots with scanning behaviors from the same source IP address observed by the darknet provides an opportunity to leverage privilege information (e.g., payload-related information) available in data from Honeypots (hence, available in the training data), but not available in data from darknet (hence, not available in the testing data). The potential value of such privilege information is illustrated in Table 5.3, which demonstrated that the performance of multi-label classification models can be enhanced by incorporating privileged information. The numbers enclosed in parentheses in Table 5.3 indicate changes to each evaluation metrics compared to the performance of the same model without using privileged information shown in Table 5.2.

The availability of privileged information at training time and its absence at test time can pose challenges. Ensuring that the model does not become overly reliant on privileged information is important to maintain its effectiveness in real-world scenarios. Future research in investigating LUPI-based approaches for amplifying threat intelligence by integrating data from Honeypots with those from darknets need to address this challenge.

Chapter 6
Learning Using Privileged Information

6.1 Background

Towards the end of the preceding chapter, it was acknowledged that the utilization of features derived from GreyNoise, which underpin the generation of threat intelligence labels, has the potential to significantly enhance the association learning. Although access to these features is limited to the training phase, the Learning Using Privileged Information (LUPI) paradigm enables the development of robust models by effectively leveraging this privileged information during training. This approach allows for the integration of additional context that can refine the learning process, thereby improving the accuracy and reliability of the resultant models. By harnessing the insights provided by GreyNoise features, the model can be trained to better discern patterns indicative of malicious activity, ultimately leading to more effective threat detection and attribution.

6.1.1 Privileged Information (PI)

Privileged information in the context of machine learning refers to additional data that is accessible during the training phase but not available during the operational phase when the model is deployed. This concept is particularly beneficial in cybersecurity applications, where such information is generally available and it can significantly enhance model performance and robustness. The paradigm that exemplifies how privileged information can be effectively utilized in machine learning is LUPI.

6.1.2 Learning Using Privileged Information (LUPI)

Learning Using Privileged Information [121], or LUPI, represents a paradigm in machine learning wherein privileged information is supplied to the learner by a teacher, in addition to the standard training data. This PI is only available for the training examples and is never available for the test examples. The goal of LUPI is to transfer knowledge from the space of privileged information to the space where the decision rule is constructed [122]. This transfer can be achieved through knowledge distillation or marginalization with weight sharing. LUPI can help to accelerate the convergence rate of learning, especially when the learning problem is hard [123].

Imagine you're training a machine learning model to classify images of different types of cars. You have a dataset of car images, but it's not very large and the images are quite similar. This makes it difficult for the model to learn to distinguish between the different car types. However, you also have access to a set of expert annotations for each image, which describe the key features of each car. This expert information is considered PI because it's not available at test time. LUPI can be used to leverage this privileged information to improve the model's performance. The model can be trained to learn from both the images and the expert annotations, and then use this knowledge to classify new images. This approach can significantly improve the model's accuracy, especially when the training data is limited.

6.1.2.1 Classical Machine Learning Paradigm

The classical paradigm of machine learning can be formally articulated as follows:

Consider a set of independent and identically distributed (iid) pairs, i.e. the training data,

$$(x_1, y_1), \ldots, (x_\ell, y_\ell), \quad x_i \in X, \quad y_i \in \{-1, +1\},$$

which are generated according to a fixed but unknown probability measure $P(x, y)$. The objective is to identify a function $y = f(x, \alpha^*)$ from a specified set of indicator functions $f(x, \alpha)$, where $\alpha \in \Lambda$, that minimizes the probability of incorrect classifications (i.e., incorrect values of $y \in \{-1, +1\}$.

In this framework, each vector $x_i \in X$ represents an example, following an unknown generator $P(x)$ for random vectors x_i. Correspondingly, $y_i \in \{-1, +1\}$ denotes its classification, defined by the conditional probability $P(y|x)$. The primary aim of the learning machine is to determine the function $y = f(x, \alpha^*)$ that ensures the lowest probability of misclassification.

Thus, the goal is to minimize the risk functional

$$R(\alpha) = \frac{1}{2} \int |y - f(x, \alpha)| dP(x, y)$$

over the set of indicator functions $f(x, \alpha)$, with $\alpha \in \Lambda$, in situations where the probability measure $P(x, y) = P(y|x)P(x)$ remains unknown, but the training data is provided.

6.1.2.2 LUPI Paradigm

The LUPI paradigm introduces a more intricate model and eliminates the need for symmetric features in training and runtime, allowing the inclusion of ancillary information in training:

Consider a collection of independent and identically distributed (iid) triplets

$$(x_1, x_1^*, y_1), \ldots, (x_\ell, x_\ell^*, y_\ell), \quad x_i \in X, \quad x_i^* \in X^*, \quad y_i \in \{-1, +1\},$$

which are generated according to a fixed but unknown probability measure $P(x, x^*, y)$. The aim is to identify, from a designated set of indicator functions $f(x, \alpha)$ with $\alpha \in \Lambda$, the function $y = f(x, \alpha^*)$ that minimizes the probability of incorrect classifications.

In the context of the LUPI paradigm, the goal remains consistent with that of the classical approach: to minimize the probability of misclassification by finding the optimal classification function within the permissible set. However, during the training phase, a richer set of information is available; specifically, triplets (x, x^*, y) are employed instead of the pairs (x, y) utilized in the classical framework. The additional data $x^* \in X^*$ is derived from a space X^*, which is, in general, different from X. For each training example (x_i, y_i), the Intelligent Teacher generates the privileged information x_i^* using some unknown conditional probability function $P(x_i^*|x_i)$.

The LUPI framework, as introduced by Vapnik and Vashist [121], extends the capabilities of Support Vector Machines (SVM) by utilizing PI to estimate slack values. The foundational concept of this initial formulation, known as SVM+, involves learning an SVM within a privileged space and determining the margin relative to this SVM for each training example. Training examples that are positioned closer to the margin are categorized as "more difficult," while those positioned further away are deemed "less difficult." Since the introduction of the new learning paradigm and the corresponding SVM+ approach, there is a growing body of work on learning with privileged information. This framework has found application across a range of challenges, including ranking [124],

clustering [125], metric learning [126], and computer vision [127, 128]. Moreover, Lapin et al. [129] demonstrated that privileged information is equivalent to weights assigned to each training example.

In the realm of cybersecurity, machine learning based detection/classification systems compares the runtime information against the known normal or anomalous states. This traditional approach relies solely on the features that are available at the runtime. In practice, many features are too expensive to collect in real-time or may be infeasible or undesirable to collect at runtime. This is a common scenario in cybersecurity where thorough analysis can generate multiple information but all these information cannot be used to build a model as they would simply be unavailable during deployment. Celik et al. [130] observed that privileged information increased precision and recall, and relatively decreased malware detection error over a system with no privileged information.

6.1.3 Impact of Privileged Information

Numerous applications illustrate a beneficial impact of privileged information (PI) on accuracy within the LUPI framework. However, this contribution may turn negative if the PI is noisy or redundant. The effect of PI is also contingent upon the size of the training set. In scenarios where the training set is notably small, the presence of PI can hinder performance, as it tends to overwhelm the learning algorithm [131]. Conversely, when the training set is of medium size, PI can enhance the learning process. In cases where the training set is large, the efficiency of learning without PI may approach that of learning with PI. Also, it is more damaging if the model becomes overreliant on the privileged information.

6.2 LUPI in Multi-Label Setting

SVM+ [121] is formulated for binary classification and has been extended for multi-class problems [132] and multi-task problems [133]. However, the application of learning using privileged information in multi-label learning settings has received limited attention in the existing literature. Wang et al. [134] employed the relationship between available information and privileged information, utilizing similarity constraints and dependencies among multiple labels captured by ranking constraints, to enhance the performance of multi-label classification in object recognition tasks. In the framework proposed by You et al. [135] for exploiting PI in multi-label setting, they exploited the dependencies

between label as privileged information i.e. for each label's learning, the other labels serve as privileged information. A multi-view multi-label model is described in [136] allows different views to serve as privileged information for each other.

Given the training data $T = \{(x_i, x_i^*, Y_i) \mid i = 1, \ldots, n\}$, where x represents the available information, x^* represents the privileged information, Y represents the target labels, and n represents the number of training instances. $Y = \{y_k \in \{-1, 1\} \mid k = 1, \ldots, q\}$ indicates the multiple labels, where q represents the number of labels.

The objective of LUPI for multi-label classification is to map the available information of an instance to its multiple labels with the help of privileged information and the label dependencies embedded in Y. Therefore, the objective function of LUPI for multi-label classification is defined as:

$$\min L = \sum_{i=1}^{n} (\ell(x_i, Y_i) + \ell^*(x_i^*, Y_i)) + C \sum_{i=1}^{n} t(x_i, Y_i) + C^* \sum_{i=1}^{n} t^*(x_i^*, Y_i) + D \sum_{i=1}^{n} p(x_i, x_i^*, Y_i)$$

where the terms $\ell(x_i, Y_i)$ and $\ell^*(x_i^*, Y_i)$ represent the loss functions of the available information classifier and the privileged information classifier, respectively. The functions $t(x_i, Y_i)$ and $t^*(x_i^*, Y_i)$ capture the dependencies among multiple labels, while $p(x_i, x_i^*, Y_i)$ reflects the constraints imposed by privileged information. The constants C, C^*, and D are the weighted parameters.

This framework represents the general approach of LUPI for multi-label classification. In this research, the decision will be made to follow established practices by adopting the maximum margin classifier as the loss function. Furthermore, the similarity between the classifier derived from available information and that obtained from privileged information will be utilized as the constraints associated with privileged information. Additionally, the ranking order of the predicted labels will serve as constraints to effectively capture multi-label dependencies.

6.3 Honeypot as Source of Privileged Information

Honeypots serve as sophisticated decoys, dynamically configured to attract malicious actors and record their activities in real-time. By simulating vulnerable systems, these traps create an enticing environment that lures attackers, enabling researchers to observe and analyze their tactics, techniques, and procedures. This dynamic setup captures payloads from attacks, which help determine the malicious intent of the attackers.

Packet payloads are not always transmitted, particularly when a scanner is focused only on identifying open ports. GreyNoise captures packet payloads when they are available, providing valuable insights. In this research, the only privileged information utilized will be the request URL in the payload, as it is sufficient to predict the labels where the multi-label classifier faced difficulties in the previous chapter. This request URL serves as a usable form of privileged information, as it can also be predicted using solely the darknet port features.

The ports scanned by GreyNoise represent another form of potential privileged information (PI). However, experiments from the previous chapter indicate that the value of the GreyNoise ports is minimal compared to the request URL within the payload. Consequently, including ports as privileged information is likely to contribute negatively, rather than positively, due to their redundancy.

6.4 Evaluation

The request URLs were systematically organized based on their shared prefixes, allowing for a structured grouping that enhances the model's ability to recognize patterns. This process involved identifying common elements within the URLs, which facilitated the subsequent application of one-hot encoding. In this research, a total of 238 unique URLs were extracted from this process, each representing a specific request that could provide valuable insights into the attack patterns being analyzed. Once the URLs were encoded, they were utilized to train the extended Support Vector Machine (SVM+) model.

To evaluate the performance of the trained multi-label classifier, the same metrics used in Chapter 5 were employed, ensuring consistency in assessment and comparison. The comparison of the results with the best model from Chapter 5 is shown in Table 6.1.

Table 6.1: Performance of classifier trained with PI over without PI.

Metrics	**Prec. Mac.**	**Rec. Mac.**	**F1 Mac.**	**Prec. Mic.**	**Rec. Mic.**	**F1 Mic.**	**Prec. Wtd.**	**Rec. Wtd.**	**F1 Wtd.**
Extended SVM+	0.88	0.87	0.87	0.85	0.87	0.86	0.84	0.83	0.83
MLWSE	0.85	0.79	0.82	0.81	0.84	0.82	0.82	0.82	0.82

While the overall performance metrics of the model may not initially suggest dramatic improvements, a more granular analysis reveals notable advancements, especially concerning individual label predictions. This nuanced examination underscores the importance of

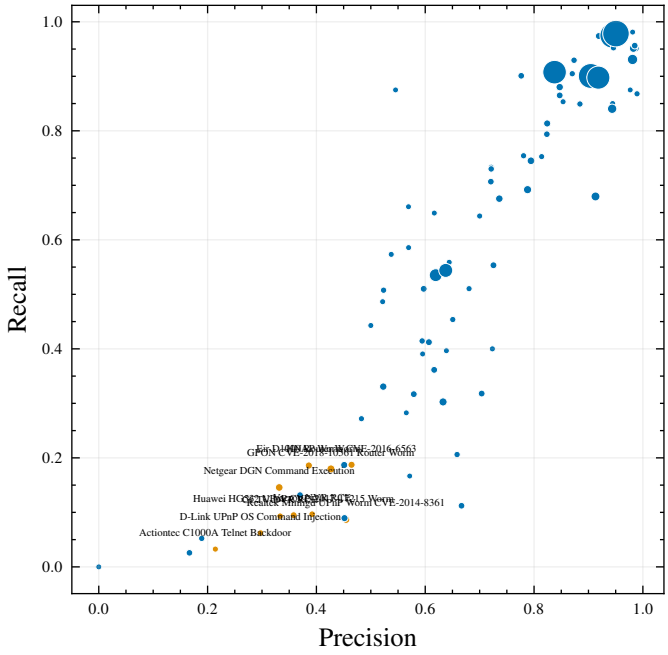
diving deeper into the data rather than relying solely on aggregate statistics. Particularly striking is the enhancement in the classification of tags related to router exploits. These tags are linked to highly specific request URLs that encapsulate distinct patterns of malicious activity. Previously, the model's reliance on darknet features alone limited its ability to accurately identify these tags, leading to suboptimal performance. However, with the integration of privileged information—specifically, the request URLs—the model has demonstrated significant gains in both precision and recall for these challenging labels. As illustrated in Figure 6.1, the model's enhanced performance in recognizing router exploit tags not only indicates an improvement in accuracy but also reflects its capacity to discern complex patterns inherent in the attack data. This shift signals that the model can now effectively identify subtle distinctions among different types of malicious behavior, which were previously overlooked.

An intriguing observation from the analysis is that the inclusion of the request URL as privileged information (PI) significantly enhanced the model's performance for certain labels, particularly those associated with router exploits and other similarly targeted categories. This improvement, however, was not universally applicable across all labels.

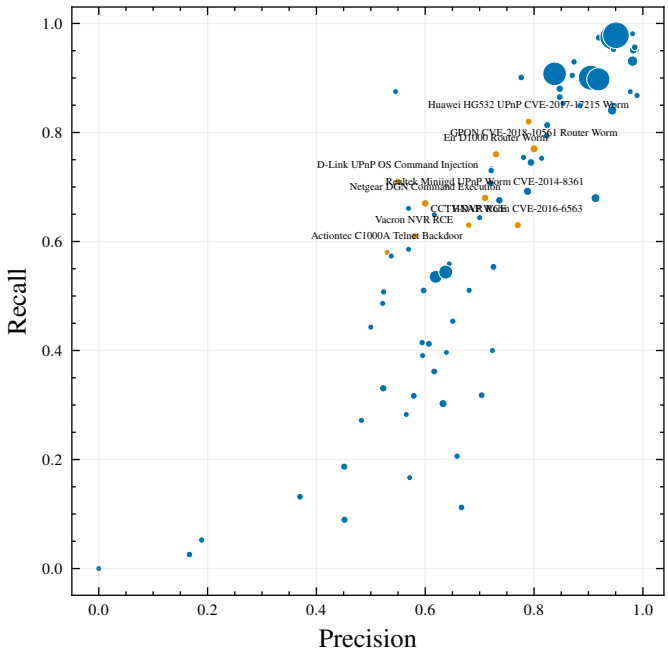
The labels that exhibited gains in precision and recall are those for which the request URLs provided clear and distinguishable characteristics within the payload data. As demonstrated in Table 6.2, it becomes evident that the specificity and clarity of the request URLs played a crucial role in the model's ability to accurately classify these particular tags. For instance, router exploit labels are often tied to very specific request patterns that reflect distinct attack methodologies, allowing the model to leverage this information effectively. Conversely, labels lacking such clear and distinctive request URLs did not experience the same level of improvement. This disparity suggests that the effectiveness of privileged information is contingent upon its relevance and applicability to the specific context of the labels being predicted.

Table 6.2: Mapping - Router Exploit Labels, Request URL, Port.

Label	**Request URL**	**Port**
Huawei H532 UPnP CVE	/ctrlt/DeviceUpgrade_1	37215
Realtek Miniigd UPnP CVE	/picsdesc.xml	52869
NETGEAR DGN Command Execution	/setup.cgi	8443



Precision-Recall plot without LUPI.



Precision-Recall plot with LUPI.

Figure 6.1: The router exploit labels were previously misclassified; however, the incorporation of privileged information has led to improved recognition of these labels.

6.5 Discussion

The findings highlight the importance of integrating sophisticated data sources into the training process, even when such information may not be accessible during runtime. It is crucial to recognize that not all additional information will contribute positively. Privileged information (PI) must offer insights that are directly relevant to the task at hand and should not be redundant in relation to the available data. If the PI fails to provide new insights or knowledge beyond what is already captured, it may lack value. Furthermore, noisy or unreliable data has the potential to degrade model performance instead of enhancing it. Information that is highly specific to the task or domain tends to be more effective as PI; for instance, specific attack patterns or user behaviors pertinent to cybersecurity are particularly valuable.

Any additional information can serve as privileged information, provided it meets several criteria: relevance, availability during training, non-redundancy, reliability, specificity, capability to guide learning, and quantifiability. The careful selection and integration of such information can significantly enhance model performance across various domains.

The insights derived from this research not only contribute to a theoretical understanding of the Learning Using Privileged Information (LUPI) framework in multi-label contexts but also offer practical implications for the enhancement of threat detection and response mechanisms. Future investigations may further refine the selection of privileged information and the incorporation of additional contextual features, thus fostering the development of more robust predictive models.

Chapter 7
Discussion and Future Work

Clustering is the most used advanced technique for decoding darknet traffic. This approach involves grouping various scanners or scanning traffic into clusters based on their distinctive features. The resultant clusters are inherently influenced by both the feature representation employed and the clustering algorithm utilized. Consequently, variations in data representation can yield different clustering outcomes, which may be acceptable when clusters are constructed with specific objectives in mind. The latent representations learned by autoencoders are typically focused on minimizing reconstruction error and remain application-agnostic, thereby making them versatile for use across various scenarios.

Many traditional clustering algorithms, such as k-means, require the specification of an input parameter indicating the desired number of clusters. Determining this parameter, commonly referred to as 'K,' can be challenging without prior examination of the data. Techniques such as t-distributed Stochastic Neighbor Embedding (t-SNE) [137] visualization and the elbow method [138] can aid in establishing an appropriate value for K. Conversely, algorithms like DBSCAN [139] can autonomously generate clusters without requiring a predefined number of clusters. Nevertheless, these methods introduce additional parameters that can significantly influence the final clustering outcomes. Advanced graph-based clustering techniques model data points as nodes within a graph, with edges representing the associations between these nodes. This framework allows for the capture of complex structures and relationships, particularly when addressing non-Euclidean spaces often encountered in network traffic analysis. The ability to represent intricate interactions among data points enhances the effectiveness of clustering in scenarios involving diverse and dynamic datasets. However, it should be noted that the results obtained from graph-based clustering methods may exhibit reduced intuitiveness compared to traditional clustering approaches.

Once clusters are obtained, interpreting these results becomes crucial. Techniques such as examining cluster centroids or employing decision trees can facilitate the interpretation of the meaning behind the clusters. Thus, the effectiveness of clustering darknet data is contingent upon multiple factors, including algorithm choice and feature representation. To enhance the analysis of clustering results, the application of optimal mass transport theory is employed to assess changes in cluster distribution over time. This methodology automates the identification of changes in cluster distributions, thus allowing alert mechanism when big changes are detected.

In addition to the utilization of the 2-Wasserstein distance, various alternative metrics for measuring the distance between probability distributions may be considered. For instance, Chen et al. [140] introduce an aggregated Wasserstein metric designed to compute the distance between two Hidden Markov Models characterized by state conditional distributions. In this approach, each clustering outcome is treated as a Gaussian distribution, thereby enabling rapid approximations of the Wasserstein metric specifically for Gaussian distributions.

Darknets undoubtedly represent one of the most effective mechanisms for capturing malicious activities on the Internet. Nevertheless, the limitations inherent in the forensic capabilities of this data—particularly the lack of insight into the intent of threat actors due to the absence of further interaction—can be substantially mitigated by establishing associations between scanning traffic data collected from darknets and other threat intelligence sources. The extensive observational capacity of darknets functions akin to an antenna, enabling not only the detection of scanning activities but also the identification of the underlying intentions driving these actions.

A particularly intriguing aspect of this integration is the opportunity for enterprises to create their own Enterprise darknet. This can be accomplished by monitoring network traffic directed at unused IP addresses or by analyzing firewall logs related to connection attempts aimed at inactive ports, following methodologies similar to those described by Richter and Berger [85]. Through this approach, detailed network-based profiles can be generated for the scanners detected within this Enterprise darknet. By mapping network characteristics to specific threat intelligence labels, critical insights can be obtained that are uniquely tailored to address threats targeting the enterprise.

Despite the promising potential of darknets, various data avenues remain unexplored within this research due to temporal constraints. One particularly valuable data source is the sequence of activities performed by scanners across different IP addresses within a large darknet. For instance, analyzing the sequence of ports scanned can provide

deeper insights into the operational patterns of malicious actors. The broad coverage afforded by darknets increases the likelihood of capturing comprehensive functionality of the scanners, thereby facilitating more accurate detection of their intent. By leveraging such multifaceted data, the understanding of malicious activities can be significantly enhanced, contributing to improved threat detection and mitigation strategies.

The utilization of privileged information presents a promising avenue for enhancing model performance, as this information can be extracted from various sources during the training phase. However, it is imperative to recognize that not all supplementary data will contribute positively to the outcomes in production environments. Careful consideration must be given to the relevance and quality of the additional information to ensure that it aligns with the objectives of the model. The integration of irrelevant or low-quality data may inadvertently introduce noise, potentially degrading performance rather than improving it.

Chapter 8
Conclusion

A substantial reservoir of information exists within darknet data, awaiting extraction through the application of suitable concepts, tools, and techniques. When harnessed effectively, this data can significantly enrich the understanding of the Internet's threat landscape, providing insights that are critical for anticipating and mitigating cyber threats.

The early detection of threats, along with the attribution of detailed threat intelligence within the complex dynamics of cyberattacks, represents a crucial yet formidable challenge in the cybersecurity field. This research introduces innovative methodologies that leverage existing technologies to tackle these challenges head-on. While certain assumptions were made to relax evaluation criteria, it is important to note that these adjustments do not undermine the validity of the results or the conclusions drawn from the analysis.

Given time constraints, the focus of this study has been specifically narrowed to understanding the darknet and a particular honeypot configuration. However, the potential for future exploration is virtually limitless, offering numerous pathways for further investigation and development in cybersecurity practices.

The analysis conducted demonstrates that the vast scale of the darknet significantly enhances the efficacy of the proposed approach in identifying multiple threat-IP associations. This methodology achieves an average lead time of one day prior to the recognition of these threats within honeypots, an advantage that is invaluable for proactive threat management. Such early insights empower cybersecurity professionals to share threat-IP associations promptly with "threat exchange" communities and other cyber-defense teams, as referenced in prior studies [141, 142].

This capability not only facilitates the swift implementation of relevant mitigation strategies but also ensures a robust and timely response to emerging cyber threats. By enabling early detection and proactive information sharing, the approach significantly

contributes to the fortification of overall cybersecurity defenses. Furthermore, it enhances collaborative efforts in threat management, underscoring the importance of a unified response in the fight against cyber adversaries. In summary, the findings of this research highlight the transformative potential of darknet data and its application in strengthening cybersecurity frameworks.

Bibliography

[1] Merit Network, Inc. (2022), "ORION: Observatory for Cyber-Risk Insights and Outages of Networks," https://www.merit.edu/initiatives/orion-network-telescope/.

[2] Durumeric, Z., E. Wustrow, and J. A. Halderman (2013) "ZMap: Fast Internet-wide Scanning and Its Security Applications," in *22nd USENIX Security Symposium (USENIX Security 13)*, pp. 605–620.

[3] Graham, R. (2022), "MASSCAN: Mass IP port scanner," https://github.com/robertdavidgraham/masscan.

[4] Cadzow, E. (2019), "Financial Impact of Mirai DDoS Attack on Dyn Revealed in New Data," https://www.corero.com/financial-impact-of-mirai-ddos-attack-on-dyn-revealed-in-new-data/.

[5] Federal Trade Commission (2022), "Equifax Data Breach Settlement," https://www.ftc.gov/enforcement/refunds/equifax-data-breach-settlement.

[6] Security Encyclopedia (2017), "NotPetya: Five Facts to Know About History's Most Destructive Cyberattack," https://www.hypr.com/security-encyclopedia/notpetya.

[7] Henriquez, M. (2022), "Five years after the WannaCry ransomware attack," https://www.securitymagazine.com/articles/97610-five-years-after-the-wannacry-ransomware-attack.

[8] Lockheed Martin (2015), "Gaining the advantage: Applying Cyber Kill Chain Methodology to Network Defense," https://www.lockheedmartin.com/content/dam/lockheed-martin/rms/documents/cyber/Gaining_the_Advantage_Cyber_Kill_Chain.pdf.

[9] Feily, M., A. Shahrestani, and S. Ramadass (2009) "A survey of botnet and botnet detection," in *2009 Third International Conference on Emerging Security Information, Systems and Technologies*, IEEE, pp. 268–273.

[10] Mishra, B. K., S. K. Srivastava, and B. K. Mishra (2014) "A quarantine model on the spreading behavior of worms in wireless sensor network," *Transaction on IoT and Cloud Computing*, **2**(1), pp. 1–12.

[11] Raftopoulos, E., E. Glatz, X. Dimitropoulos, and A. Dainotti (2015) "How dangerous is internet scanning? a measurement study of the aftermath of an internet-wide scan," in *Traffic Monitoring and Analysis: 7th International Workshop, TMA 2015, Barcelona, Spain, April 21-24, 2015. Proceedings 7*, Springer, pp. 158–172.

[12] Pang, R., V. Yegneswaran, P. Barford, V. Paxson, and L. Peterson (2004) "Characteristics of internet background radiation," in *Proceedings of the 4th ACM SIGCOMM conference on Internet measurement*, pp. 27–40.

[13] Wustrow, E., M. Karir, M. Bailey, F. Jahanian, and G. Huston (2010) "Internet background radiation revisited," in *Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement*, pp. 62–74.

[14] Malécot, E. L. and D. Inoue (2013) "The carna botnet through the lens of a network telescope," in *International Symposium on Foundations and Practice of Security*, Springer, pp. 426–441.

[15] Dagon, D., C. C. Zou, and W. Lee (2006) "Modeling Botnet Propagation Using Time Zones." in *NDSS*, vol. 6, pp. 2–13.

[16] Gu, G., P. A. Porras, V. Yegneswaran, M. W. Fong, and W. Lee (2007) "Bothunter: Detecting malware infection through ids-driven dialog correlation." in *USENIX Security Symposium*, vol. 7, pp. 1–16.

[17] Moore, D., C. Shannon, and K. Claffy (2002) "Code-Red: a case study on the spread and victims of an Internet worm," in *Proceedings of the 2nd ACM SIGCOMM Workshop on Internet measurment*, pp. 273–284.

[18] Moore, D., V. Paxson, S. Savage, C. Shannon, S. Staniford, and N. Weaver (2003) "Inside the slammer worm," *IEEE Security & Privacy*, **1**(4), pp. 33–39.

[19] Moore, D. (2003), "The spread of the sapphire/slammer worm," `https://www.caida.org/catalog/papers/2003_sapphire/`.

[20] Harder, U., M. W. Johnson, J. T. Bradley, and W. J. Knottenbelt (2006) "Observing internet worm and virus attacks with a small network telescope," *Electronic Notes in Theoretical Computer Science*, **151**(3), pp. 47–59.

[21] Irwin, B. (2012) "A network telescope perspective of the Conficker outbreak," in *2012 Information Security for South Africa*, IEEE, pp. 1–8.

[22] Bailey, M., E. Cooke, F. Jahanian, A. Myrick, and S. Sinha (2006) "Practical darknet measurement," in *2006 40th Annual Conference on Information Sciences and Systems*, IEEE, pp. 1496–1501.

[23] Dainotti, A., R. Amman, E. Aben, and K. C. Claffy (2012) "Extracting benefit from harm: using malware pollution to analyze the impact of political and geophysical events on the Internet," *ACM SIGCOMM Computer Communication Review*, **42**(1), pp. 31–39.

[24] Dainotti, A., C. Squarcella, E. Aben, K. C. Claffy, M. Chiesa, M. Russo, and A. Pescapé (2011) "Analysis of country-wide internet outages caused by censorship," in *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*, pp. 1–18.

[25] Inoue, D., K. Yoshioka, M. Eto, Y. Hoshizawa, and K. Nakao (2008) "Malware behavior analysis in isolated miniature network for revealing malware's network activity," in *2008 IEEE International Conference on Communications*, IEEE, pp. 1715–1721.

[26] Dacier, M., C. Leita, O. Thonnard, H. V. Pham, and E. Kirda (2010) "Assessing cybercrime through the eyes of the WOMBAT," in *Cyber Situational Awareness*, Springer, pp. 103–136.

[27] Yegneswaran, V., P. Barford, and J. Ullrich (2003) "Internet intrusions: Global characteristics and prevalence," *ACM SIGMETRICS Performance Evaluation Review*, **31**(1), pp. 138–147.

[28] Thonnard, O. and M. Dacier (2008) "Actionable knowledge discovery for threats intelligence support using a multi-dimensional data mining methodology," in *2008 ieee international conference on data mining workshops*, IEEE, pp. 154–163.

[29] Fachkha, C., E. Bou-Harb, A. Boukhtouta, S. Dinh, F. Iqbal, and M. Debbabi (2012) "Investigating the dark cyberspace: Profiling, threat-based analysis and correlation," in *2012 7th International Conference on Risks and Security of Internet and Systems (CRiSIS)*, IEEE, pp. 1–8.

[30] Panjwani, S., S. Tan, K. M. Jarrin, and M. Cukier (2005) "An experimental evaluation to determine if port scans are precursors to an attack," in *2005 International Conference on Dependable Systems and Networks (DSN'05)*, IEEE, pp. 602–611.

[31] Bou-Harb, E., M. Husák, M. Debbabi, and C. Assi (2017) "Big data sanitization and cyber situational awareness: a network telescope perspective," *IEEE transactions on big data*, **5**(4), pp. 439–453.

[32] Eto, M., D. Inoue, M. Suzuki, and K. Nakao (2009) "A statistical packet inspection for extraction of spoofed IP packets on darknet," in *Proceedings of the Joint Workshop on Information Security, Kaohsiung, Taiwan.*

[33] Ohta, M., Y. Kanda, K. Fukuda, and T. Sugawara (2011) "Analysis of spoofed IP traffic using time-to-live and identification fields in IP headers," in *2011*

IEEE Workshops of International Conference on Advanced Information Networking and Applications, IEEE, pp. 355–361.

[34] Bi, J., P. Hu, and P. Li (2010) "Study on classification and characteristics of source address spoofing attacks in the internet," in *2010 Ninth International Conference on Networks*, IEEE, pp. 226–230.

[35] Shannon, C. and D. Moore (2004) "The spread of the witty worm," *IEEE Security & Privacy*, **2**(4), pp. 46–50.

[36] Antonakakis, M., T. April, M. Bailey, M. Bernhard, E. Bursztein, J. Cochran, Z. Durumeric, J. A. Halderman, L. Invernizzi, M. Kallitsis, et al. (2017) "Understanding the mirai botnet," in *26th USENIX security symposium (USENIX Security 17)*, pp. 1093–1110.

[37] Torabi, S., E. Bou-Harb, C. Assi, M. Galluscio, A. Boukhtouta, and M. Debbabi (2018) "Inferring, characterizing, and investigating internet-scale malicious iot device activities: A network telescope perspective," in *2018 48th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, IEEE, pp. 562–573.

[38] Cabana, O., A. M. Youssef, M. Debbabi, B. Lebel, M. Kassouf, R. Atallah, and B. L. Agba (2021) "Threat intelligence generation using network telescope data for industrial control systems," *IEEE Transactions on Information Forensics and Security*, **16**, pp. 3355–3370.

[39] Zeng, P., G. Lin, L. Pan, Y. Tai, and J. Zhang (2020) "Software vulnerability analysis and discovery using deep learning techniques: A survey," *IEEE Access*, **8**, pp. 197158–197172.

[40] Pour, M. S., E. Bou-Harb, K. Varma, N. Neshenko, D. A. Pados, and K.-K. R. Choo (2019) "Comprehending the IoT cyber threat landscape: A data dimensionality reduction technique to infer and characterize Internet-scale IoT probing campaigns," *Digital Investigation*, **28**, pp. S40–S49.

[41] Sarabi, A. and M. Liu (2018) "Characterizing the internet host population using deep learning: A universal and lightweight numerical embedding," in *Proceedings of the Internet Measurement Conference 2018*, pp. 133–146.

[42] Kallitsis, M., R. Prajapati, V. Honavar, D. Wu, and J. Yen (2022) "Detecting and Interpreting Changes in Scanning Behavior in Large Network Telescopes," *IEEE Transactions on Information Forensics and Security*, **17**, pp. 3611–3625.

[43] Jin, Y., Z.-L. Zhang, K. Xu, F. Cao, and S. Sahu (2007) "Identifying and tracking suspicious activities through IP gray space analysis," in *Proceedings of the 3rd Annual ACM Workshop on Mining Network Data*, pp. 7–12.

[44] Li, Z., A. Goyal, Y. Chen, and V. Paxson (2009) "Automating analysis of large-scale botnet probing events," in *Proceedings of the 4th International Symposium on Information, Computer, and Communications Security*, pp. 11–22.

[45] ——— (2010) "Towards situational awareness of large-scale botnet probing events," *IEEE Transactions on Information Forensics and Security*, **6**(1), pp. 175–188.

[46] Akiyoshi, R., D. Kotani, and Y. Okabe (2018) "Detecting emerging large-scale vulnerability scanning activities by correlating low-interaction honeypots with darknet," in *2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC)*, vol. 2, IEEE, pp. 658–663.

[47] Metongnon, L. and R. Sadre (2019) "Prevalence of IoT protocols in telescope and honeypot measurements," *Journal of Cyber Security and Mobility*, pp. 321–340.

[48] Shaikh, F., E. Bou-Harb, N. Neshenko, A. P. Wright, and N. Ghani (2018) "Internet of malicious things: Correlating active and passive measurements for inferring and characterizing internet-scale unsolicited iot devices," *IEEE Communications Magazine*, **56**(9), pp. 170–177.

[49] Sharma, S. (2022), "Ransomware is top cyberattack type, as manufacturing gets hit hardest," `https://www.csoonline.com/article/3651489/ransomware-is-top-cyberattack-type-as-manufacturing-gets-hit-hardest.html`.

[50] Nguyen, T. T. and G. Armitage (2008) "A survey of techniques for internet traffic classification using machine learning," *IEEE communications surveys & tutorials*, **10**(4), pp. 56–76.

[51] Shon, T. and J. Moon (2007) "A hybrid machine learning approach to network anomaly detection," *Information Sciences*, **177**(18), pp. 3799–3821.

[52] Abbasi, M., A. Shahraki, and A. Taherkordi (2021) "Deep learning for network traffic monitoring and analysis (NTMA): A survey," *Computer Communications*, **170**, pp. 19–41.

[53] Durumeric, Z., M. Bailey, and J. A. Halderman (2014) "An Internet-wide View of Internet-wide Scanning," in *23rd USENIX Security Symposium (USENIX Security 14)*, pp. 65–78.

[54] Moore, D., C. Shannon, G. M. Voelker, and S. Savage, "Network telescopes: Technical report," `https://www.cs.unc.edu/~jeffay/courses/nidsS05/measurement/moore-telescopes04.pdf`.

[55] MaxMind, "GeoIP products of MaxMind," `https://dev.maxmind.com/geoip/geoip2/geolite2/`.

[56] CAIDA, "Routeviews Prefix to AS mappings Dataset (pfx2as) for IPv4 and IPv6," `https://www.caida.org/catalog/datasets/routeviews-prefix2as/`.

[57] The Censys Team (2017), "Censys.io," https://censys.io.

[58] Siby, S. (2014), "Default TTL (Time To Live) Values of Different OS," https://subinsb.com/default-device-ttl-values.

[59] Hiesgen, R., M. Nawrocki, A. King, A. Dainotti, T. Schmidt, and M. Wählisch (2022) "Spoki: Unveiling a New Wave of Scanners through a Reactive Network Telescope," in *31st USENIX Security Symposium (USENIX Security 22)*, USENIX Association, Boston, MA.

[60] Benson, K., A. Dainotti, k. claffy, A. C. Snoeren, and M. Kallitsis (2015) "Leveraging Internet Background Radiation for Opportunistic Network Analysis," in *Proceedings of the 2015 ACM Conference on Internet Measurement Conference*, IMC '15.

[61] Bengio, Y., A. Courville, and P. Vincent (2013) "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, **35**(8), pp. 1798–1828.

[62] Abdi, H. and L. J. Williams (2010) "Principal component analysis," *Wiley interdisciplinary reviews: computational statistics*, **2**(4), pp. 433–459.

[63] Fukuda, K., T. Hirotsu, O. Akashi, and T. Sugawara (2010) "A pca analysis of daily unwanted traffic," in *2010 24th IEEE International Conference on Advanced Information Networking and Applications*, IEEE, pp. 377–384.

[64] Ring, M., A. Dallmann, D. Landes, and A. Hotho (2017) "IP2Vec: Learning similarities between ip addresses," in *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, IEEE, pp. 657–666.

[65] Mikolov, T., K. Chen, G. Corrado, and J. Dean (2013) "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*.

[66] Cohen, D., Y. Mirsky, M. Kamp, T. Martin, Y. Elovici, R. Puzis, and A. Shabtai (2020) "DANTE: A framework for mining and monitoring darknet traffic," in *European Symposium on Research in Computer Security*, Springer, pp. 88–109.

[67] Gioacchini, L., L. Vassio, M. Mellia, I. Drago, Z. B. Houidi, and D. Rossi (2021) "DarkVec: automatic analysis of darknet traffic with word embeddings," in *Proceedings of the 17th International Conference on emerging Networking EXperiments and Technologies*, pp. 76–89.

[68] Hinton, G. and R. Salakhutdinov (2006) "Reducing the Dimensionality of Data with Neural Networks," *Science NY*, **313**, pp. 504–7.

[69] Yang, B., X. Fu, N. D. Sidiropoulos, and M. Hong (2017) "Towards k-means-friendly spaces: Simultaneous deep learning and clustering," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, JMLR. org, pp. 3861–3870.

[70] Kingma, D. P. and M. Welling (2013) "Auto-Encoding Variational Bayes," *arXiv e-prints*, arXiv:1312.6114, `1312.6114`.

[71] Downey, A. B. (2008) "A novel changepoint detection algorithm," *arXiv preprint arXiv:0812.1237*.

[72] Page, E. S. (1954) "Continuous inspection schemes," *Biometrika*, **41**(1/2), pp. 100–115.

[73] Wang, H., D. Zhang, and K. G. Shin (2002) "Detecting SYN flooding attacks," in *Proceedings. Twenty-first annual joint conference of the IEEE computer and communications societies*, vol. 3, IEEE, pp. 1530–1539.

[74] Siris, V. A. and F. Papagalou (2004) "Application of anomaly detection algorithms for detecting SYN flooding attacks," in *IEEE Global Telecommunications Conference, 2004. GLOBECOM'04.*, vol. 4, IEEE, pp. 2050–2054.

[75] Bo, C., B. X. Fang, and X. C. Yun (2005) "A new approach for early detection of internet worms based on connection degree," in *2005 International Conference on Machine Learning and Cybernetics*, vol. 4, IEEE, pp. 2424–2430.

[76] Chan, J., C. Leckie, and T. Peng (2006) "Hitlist worm detection using source IP address history," in *Proceedings of Australian Telecommunication Networks and Applications Conference*.

[77] Ahmed, E., A. Clark, and G. Mohay (2008) "A novel sliding window based change detection algorithm for asymmetric traffic," in *2008 IFIP International Conference on Network and Parallel Computing*, IEEE, pp. 168–175.

[78] ——— (2009) "Effective change detection in large repositories of unsolicited traffic," in *2009 Fourth International Conference on Internet Monitoring and Protection*, IEEE, pp. 1–6.

[79] Inoue, D., M. Eto, K. Yoshioka, S. Baba, K. Suzuki, J. Nakazato, K. Ohtaka, and K. Nakao (2008) "nicter: An incident analysis system toward binding network monitoring with malware analysis," in *2008 WOMBAT Workshop on Information Security Threats Data Collection and Sharing*, IEEE, pp. 58–66.

[80] Sun, S., C. Zhang, and G. Yu (2006) "A Bayesian network approach to traffic flow forecasting," *IEEE Transactions on intelligent transportation systems*, **7**(1), pp. 124–132.

[81] Kolouri, S., S. R. Park, M. Thorpe, D. Slepcev, and G. K. Rohde (2017) "Optimal mass transport: Signal processing and machine-learning applications," *IEEE signal processing magazine*, **34**(4), pp. 43–59.

[82] Inoue, D., K. Yoshioka, M. Eto, M. Yamagata, E. Nishino, J. Takeuchi, K. Ohkouchi, and K. Nakao (2008) "An incident analysis system NICTER and its analysis engines based on data mining techniques," in *International Conference on Neural Information Processing*, Springer, pp. 579–586.

[83] Bou-Harb, E., M. Debbabi, and C. Assi (2013) "A systematic approach for detecting and clustering distributed cyber scanning," *Computer Networks*, **57**(18), pp. 3826–3839.

[84] Soro, F., M. Allegretta, M. Mellia, I. Drago, and L. M. Bertholdo (2020) "Sensing the noise: Uncovering communities in darknet traffic," in *2020 Mediterranean Communication and Computer Networking Conference (MedComNet)*, IEEE, pp. 1–8.

[85] Richter, P. and A. Berger (2019) "Scanning the Scanners: Sensing the Internet from a Massively Distributed Network Telescope," in *Proceedings of ACM IMC 2019*, Amsterdam, Netherlands.

[86] Cowie, B. and B. Irwin (2010) "Data classification for artificial intelligence construct training to aid in network incident identification using network telescope data," in *Proceedings of the 2010 Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists*, pp. 356–360.

[87] Højsgaard, S., D. Edwards, and S. Lauritzen (2012) *Graphical Models with R*, Springer, Boston, MA, USA.

[88] Russell, S. and P. Norvig (2020) *Artificial Intelligence: A Modern Approach, 4th ed.*, Pearson, USA.

[89] Rossow, C. (2014) "Amplification Hell: Revisiting Network Protocols for DDoS Abuse," in *Proceedings of the 2014 Network and Distributed System Security (NDSS) Symposium*.

[90] Sahota, J. and N. Vlajic (2021) "Mozi IoT malware and its botnets: From theory to real-world observations," in *2021 International Conference on Computational Science and Computational Intelligence (CSCI)*, IEEE, pp. 698–703.

[91] Alrawi, O., C. Lever, K. Valakuzhy, R. Court, K. Snow, F. Monrose, and M. Antonakakis (2021) "The Circle Of Life: A Large-Scale Study of The IoT Malware Lifecycle," in *30th USENIX Security Symposium (USENIX Security 21)*, USENIX Association, pp. 3505–3522.

[92] Young, C., R. McArdle, N. A. Le Khac, and K. K. R. Choo (2020) *Forensic Investigation of Ransomware Activities—Part 1*, Springer International Publishing, Cham, pp. 51–77.

[93] Akbanov, M., V. G. Vassilakis, and M. D. Logothetis (2019) "Ransomware detection and mitigation using software-defined networking: The case of WannaCry," *Computers & Electrical Engineering*, **76**, pp. 111–121.

[94] Shodan (2009), "Shodan Search Engine," `www.shodan.io`.

[95] Collins, M. (2021), "Acknowledged Scanners (Version 1.0)," `https://gitlab.com/mcollins_at_isi/acknowledged_scanners`.

[96] Normshield (2020), "Mature your TPRM Program with NormShield's FAIR-based Financial Impact," `https://blackkite.com/wp-content/uploads/2020/09/FAIR_Report_2.pdf`.

[97] Jonker, M., A. King, J. Krupp, C. Rossow, A. Sperotto, and A. Dainotti (2017) "Millions of targets under attack: a macroscopic characterization of the DoS ecosystem," in *Proceedings of the 2017 Internet Measurement Conference*, pp. 100–113.

[98] Fachkha, C. and M. Debbabi (2015) "Darknet as a source of cyber intelligence: Survey, taxonomy, and characterization," *IEEE Communications Surveys & Tutorials*, **18**(2), pp. 1197–1227.

[99] GreyNoise (2022), "GreyNoise: GreyNoise is THE source for understanding internet noise," `https://www.greynoise.io/`.

[100] Ban, T., M. Eto, S. Guo, D. Inoue, K. Nakao, and R. Huang (2015) "A study on association rule mining of darknet big data," in *2015 International Joint Conference on Neural Networks (IJCNN)*, IEEE, pp. 1–7.

[101] Zhang, M. L. and Z. H. Zhou (2013) "A review on multi-label learning algorithms," *IEEE transactions on knowledge and data engineering*, **26**(8), pp. 1819–1837.

[102] Nguyen, T. T., M. T. Dang, A. V. Luong, A. W. C. Liew, T. Liang, and J. McCall (2019) "Multi-label classification via incremental clustering on an evolving data stream," *Pattern Recognition*, **95**, pp. 96–113.

[103] Charte, F., A. Rivera, M. J. d. Jesus, and F. Herrera (2014) "Concurrence among imbalanced labels and its influence on multilabel resampling algorithms," in *International conference on hybrid artificial intelligence systems*, Springer, pp. 110–121.

[104] Charte, F., A. J. Rivera, M. J. del Jesus, and F. Herrera (2015) "MLSMOTE: Approaching imbalanced multilabel learning through synthetic instance generation," *Knowledge-Based Systems*, **89**, pp. 385–397.

[105] Tsoumakas, G. and I. Katakis (2007) "Multi-label classification: An overview," *International Journal of Data Warehousing and Mining (IJDWM)*, **3**(3), pp. 1–13.

[106] Godbole, S. and S. Sarawagi (2004) "Discriminative methods for multi-labeled classification," in *Pacific-Asia conference on knowledge discovery and data mining*, Springer, pp. 22–30.

[107] Bhatia, K., H. Jain, P. Kar, M. Varma, and P. Jain (2015) "Sparse local embeddings for extreme multi-label classification," *Advances in neural information processing systems*, **28**.

[108] Read, J., B. Pfahringer, G. Holmes, and E. Frank (2009) "Classifier chains for multi-label classification," in *Joint European conference on machine learning and knowledge discovery in databases*, Springer, pp. 254–269.

[109] Tsoumakas, G. and I. Vlahavas (2007) "Random k-labelsets: An ensemble method for multilabel classification," in *European conference on machine learning*, Springer, pp. 406–417.

[110] Xia, Y., K. Chen, and Y. Yang (2021) "Multi-label classification with weighted classifier selection and stacked ensemble," *Information Sciences*, **557**, pp. 421–442.

[111] Jasinska-Kobus, K., M. Wydmuch, K. Dembczynski, M. Kuznetsov, and R. Busa-Fekete (2020) "Probabilistic label trees for extreme multi-label classification," *arXiv preprint arXiv:2009.11218*.

[112] Babbar, R. and B. Schölkopf (2019) "Data scarcity, robustness and extreme multi-label classification," *Machine Learning*, **108**(8), pp. 1329–1351.

[113] Bertsimas, D. and J. Dunn (2017) "Optimal classification trees," *Machine Learning*, **106**, pp. 1039–1082.

[114] Atch, D., G. Regev, and R. Bevington (2021), "How to proactively defend against Mozi IoT botnet," `https://www.microsoft.com/en-us/security/blog/2021/08/19/how-to-proactively-defend-against-mozi-iot-botnet/`.

[115] Widmer, G. and M. Kubat (1996) "Learning in the presence of concept drift and hidden contexts," *Machine learning*, **23**, pp. 69–101.

[116] Andresini, G., F. Pendlebury, F. Pierazzi, C. Loglisci, A. Appice, and L. Cavallaro (2021) "Insomnia: Towards concept-drift robustness in network intrusion detection," in *Proceedings of the 14th ACM workshop on artificial intelligence and security*, pp. 111–122.

[117] Polikar, R. (2012) "Ensemble learning," *Ensemble machine learning: Methods and applications*, pp. 1–34.

[118] Zhao, K., T. Matsukawa, and E. Suzuki (2018) "Retraining: A simple way to improve the ensemble accuracy of deep neural networks for image classification," in *2018 24th international conference on pattern recognition (ICPR)*, IEEE, pp. 860–867.

[119] Sun, L., P. Ye, G. Lyu, S. Feng, G. Dai, and H. Zhang (2020) "Weakly-supervised multi-label learning with noisy features and incomplete labels," *Neurocomputing*, **413**, pp. 61–71.

[120] Teisseyre, P. (2021) "Classifier chains for positive unlabelled multi-label learning," *Knowledge-Based Systems*, **213**, p. 106709.

[121] Vapnik, V. and A. Vashist (2009) "A new learning paradigm: Learning using privileged information," *Neural networks*, **22**(5-6), pp. 544–557.

[122] Pechyony, D. and V. Vapnik (2010) "On the theory of learnining with privileged information," *Advances in neural information processing systems*, **23**.

[123] Vapnik, V., R. Izmailov, et al. (2015) "Learning using privileged information: similarity control and knowledge transfer." *J. Mach. Learn. Res.*, **16**(1), pp. 2023–2049.

[124] Sharmanska, V., N. Quadrianto, and C. H. Lampert (2013) "Learning to rank using privileged information," in *Proceedings of the IEEE international conference on computer vision*, pp. 825–832.

[125] Feyereisl, J. and U. Aickelin (2012) "Privileged information for data clustering," *Information Sciences*, **194**, pp. 4–23.

[126] Fouad, S., P. Tino, S. Raychaudhury, and P. Schneider (2013) "Incorporating privileged information through metric learning," *IEEE transactions on neural networks and learning systems*, **24**(7), pp. 1086–1098.

[127] Feyereisl, J., S. Kwak, J. Son, and B. Han (2014) "Object localization based on structural SVM using privileged information," *Advances in Neural Information Processing Systems*, **27**.

[128] Sharmanska, V. and N. Quadrianto (2017) "In the era of deep convolutional features: Are attributes still useful privileged data?" *Visual Attributes*, pp. 31–48.

[129] Lapin, M., M. Hein, and B. Schiele (2014) "Learning using privileged information: SVM+ and weighted SVM," *Neural Networks*, **53**, pp. 95–108.

[130] Celik, Z. B., P. McDaniel, R. Izmailov, N. Papernot, R. Sheatsley, R. Alvarez, and A. Swami (2018) "Detection under privileged information," in *Proceedings of the 2018 on Asia Conference on Computer and Communications Security*, pp. 199–206.

[131] Nouretdinov, I. (2022) "On Efficiency of Learning Under Privileged Information," in *Conformal and Probabilistic Prediction with Applications*, PMLR, pp. 239–252.

[132] Liu, J., W. Zhu, and P. Zhong (2013) "A new multi-class support vector algorithm based on privileged information," *Journal of Information and Computational Science*, **2**.

[133] Ji, Y., S. Sun, and Y. Lu (2012) "Multitask multiclass privileged information support vector machines," in *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, IEEE, pp. 2323–2326.

[134] Wang, S., S. Chen, T. Chen, and X. Shi (2018) "Learning with privileged information for multi-label classification," *Pattern Recognition*, **81**, pp. 60–70.

[135] You, S., C. Xu, Y. Wang, C. Xu, and D. Tao (2017) "Privileged multi-label learning," *arXiv preprint arXiv:1701.07194*.

[136] Xiao, Y., J. Chen, B. Liu, L. Zhao, X. Kong, and Z. Hao (2024) "A new multi-view multi-label model with privileged information learning," *Information Sciences*, **656**, p. 119911.

[137] van der Maaten, L. and G. Hinton (2008) "Visualizing Data using t-SNE," *Journal of Machine Learning Research*, **9**(86), pp. 2579–2605.

[138] Syakur, M. A., B. K. Khotimah, E. Rochman, and B. D. Satoto (2018) "Integration k-means clustering method and elbow method for identification of the best customer profile cluster," in *IOP conference series: materials science and engineering*, vol. 336, IOP Publishing, p. 012017.

[139] Khan, K., S. U. Rehman, K. Aziz, S. Fong, and S. Sarasvady (2014) "DBSCAN: Past, present and future," in *The fifth international conference on the applications of digital information and web technologies (ICADIWT 2014)*, IEEE, pp. 232–238.

[140] Chen, Y., J. Ye, and J. Li (2016) "A distance for HMMS based on aggregated Wasserstein metric and state registration," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14*, Springer, pp. 451–466.

[141] Thomas, K., R. Amira, A. Ben-Yoash, O. Folger, A. Hardon, A. Berger, E. Bursztein, and M. Bailey (2016) "The Abuse Sharing Economy: Understanding the Limits of Threat Exchanges," in *Research in Attacks, Intrusions, and*

Defenses (F. Monrose, M. Dacier, G. Blanc, and J. Garcia-Alfaro, eds.), Springer International Publishing, Cham, pp. 143–164.

[142] Li, V. G., M. Dunn, P. Pearce, D. McCoy, G. M. Voelker, and S. Savage (2019) "Reading the tea leaves: A comparative analysis of threat intelligence," in *28th USENIX security symposium (USENIX Security 19)*, pp. 851–867.

Vita

Rupesh Prajapati

Rupesh Prajapati completed his Bachelor in Engineering (B.E.) in Electronics and Communication from Institute of Engineering (IOE), Tribhuvan University, Nepal, in November 2015. Thereafter, he worked in industry for one and half years (2015-2017) in LIS Nepal Private Limited as a software developer. Thereafter, he joined the PhD program in Informatics at The Pennsylvania State University in 2017. During his PhD, he conducted research at the intersection of cybersecurity and artificial intelligence (AI), focusing on how advancements in AI can be utilized to extract and harness threat intelligence from darknet data.

Selected Publications

1. Prajapati, Rupesh, Vasant Honavar, Dinghao Wu and John Yen. Shedding light into the darknet: scanning characterization and detection of temporal changes, CoNEXT '21: Proceedings of the 17th International Conference on emerging Networking EXperiments and Technologies.

2. Prajapati, Rupesh, Kallitsis, Michalis, Vasant Honavar, Dinghao Wu and John Yen. Coupling Network Telescope and Interactive Honeypot Data to Enhance Threat Intelligence via Machine Learning, Submitted to IEEE Transactions of Information Forensics and Security. 2025.

3. Kallitsis, Michalis, Rupesh Prajapati, Vasant Honavar, Dinghao Wu, and John Yen. Detecting and Interpreting Changes in Scanning Behavior in Large Network Telescopes. In IEEE Transactions of Information Forensics and Security. 2022.

4. Liu, Xiao, Xiaoting Li, Rupesh Prajapati and Dinghao Wu. DeepFuzz: Automatic Generation of Syntax Valid C Programs for Fuzz Testing, 33rd AAAI Conference on Artificial Intelligence (AAAI 2019).

5. Li, Xiaoting, Xiao Liu, Lingwei Chen, Rupesh Prajapati and Dinghao Wu. ALPHAPROG: Reinforcement Generation of Valid Programs for Compiler Fuzzing, 34th Annual Conference on Innovative Applications of Artificial Intelligence (IAAI-22).

6. Li, Xiaoting, Xiao Liu, Lingwei Chen, Rupesh Prajapati and Dinghao Wu. FuzzBoost: Reinforcement Compiler Fuzzing (ICICS-22).