Automated IT-OT Device Classification

Joshua Dsouza, Shubham Nayak, Ojas Taskar, Parth Pawar, Saniya Deshmukh
E&TC Batch ‘24

Abstract

This research introduces a novel algorithm for automated classification of network devices into Information Technology (IT) and Operational Technology (OT) categories. The proposed algorithm employs a comprehensive approach, involving data preprocessing, payload tokenization, feature scoring, and machine learning. It leverages packet source port numbers, protocol names, and payload content to calculate IT and OT scores for each packet. Additionally, the algorithm identifies email, web-address, and hexadecimal patterns in payload data to enhance classification accuracy. The resulting dataset is used for model training, followed by evaluation using a Random Forest Classifier. Experimental results demonstrate the effectiveness of the proposed algorithm in accurately categorizing network devices, showcasing its potential for streamlining IT/OT classification in cybersecurity applications.

Introduction

In the era of interconnected systems and digital networks, the demarcation between Information Technology (IT) and Operational Technology (OT) has become increasingly critical for effective cybersecurity management. The coexistence of IT, which encompasses traditional computing and communication technologies, and OT, which pertains to industrial control and automation systems, has resulted in complex challenges in safeguarding critical infrastructures. Effective identification and differentiation of IT and OT devices are essential for targeted security measures, incident response, and risk mitigation.

Manual classification of network devices into IT and OT categories can be time-consuming and error-prone, especially given the diverse communication protocols, port numbers, and payload patterns that characterize each domain. This paper addresses this challenge by proposing an algorithmic approach for automated IT/OT classification. The algorithm amalgamates advanced techniques from natural language processing and machine learning to discern key attributes of network traffic and categorize devices accordingly.

The algorithm commences with data preprocessing, including the creation of datasets containing port numbers, protocol names, and payload content for both IT and OT network devices. Payloads are tokenized, and a Word2Vec model is employed to capture semantic relationships between tokens. A scoring mechanism is introduced, evaluating factors such as port numbers, protocol names, and payload keywords, enhancing the algorithm's capability to distinguish between IT and OT devices. Moreover, the algorithm leverages the unique patterns present in payload data. Email and web page patterns are detected to identify IT-related communications, while hexadecimal data patterns are employed to discern OT communication

in payload content. These additions to the scoring process contribute to a nuanced classification of network devices.
To validate the algorithm's effectiveness, a Random Forest Classifier is trained on a dataset generated using the algorithm's scoring outcomes. The resulting model is evaluated on testing data, yielding insights into its performance and classification accuracy. The proposed algorithm has the potential to streamline the identification of IT and OT devices, thus fortifying cybersecurity practices. This research not only presents a comprehensive algorithmic framework but also contributes to the ongoing discourse on the automated classification of heterogeneous network devices, presenting a promising stride towards improved cyber threat detection and mitigation in modern networked environments.

Problem Statement

The convergence of Information Technology (IT) and Operational Technology (OT) has introduced a critical challenge in modern cybersecurity—accurate and automated classification of network devices into their respective domains. The absence of a systematic and efficient method for distinguishing between IT and OT devices hinders targeted security measures, incident response, and effective risk management. Manual classification processes are time-consuming, error-prone, and struggle to accommodate the diverse range of communication protocols, port numbers, and payload patterns inherent to IT and OT domains. As a consequence, there exists a pressing need for an algorithmic solution capable of automating the differentiation between IT and OT devices, thereby enhancing the precision of cybersecurity strategies and bolstering the protection of critical infrastructures. This research addresses this challenge by proposing an innovative algorithm that harnesses the power of natural language processing, pattern recognition, and machine learning techniques to achieve accurate and efficient IT/OT classification in dynamic network environments.

Literature Review

1. Comparative study of various databases including p0f, grassmarlin, network miner and Sartori dataset (https://drive.google.com/file/d/1YW5JJ0hfbXdMgQMKyD0xjASTqXp7MhsZ/view?usp=sharing)

2. Listing relevant features for OT and IT systems and tockenising them (https://drive.google.com/file/d/1t498MSy_BlkpsGt1zxTS8EUs9s0Tlpe1/view?usp=sharing)

3. Study on function codes and banner grabbing mechanisms. For Modbus - (https://drive.google.com/file/d/1AGHZu1hreWdqqGFv60RKFQpJJQglxfQx/view?usp=sharing)

4. Utilization of protocols and payload values to generate an IT-OT classification system (https://docs.google.com/document/d/1DetFjFWYrCQIO0DG6AN2mkTnm_Wq0Yig)

5. Generation of a compiled database for mac vendor correlation utilizing various resources

6. Comparative study and implementation of various MAC algorithms in case of absence of direct match (docs.google.com/document/d/1rhOg-9YL_YDja4a7yjtmnupY66xxQrD41FY_D8pZ2Vs/0)

7. Studied the previously accomplished work and identified the possible challenges (docs.google.com/presentation/d/171veh3YhEoAX8DmlEqvILj21uk_bIwI2)

8. Possible methods for identification of the purdue level of a device by utilizing the PCAP
    - Protocol and Port number
    - Metadata in PCAP
    - Identifying function via asset correlation

9. The Google Collab link for the experiment and its working flow is attached below : (https://docs.google.com/document/d/1DetFjFWYrCQIO0DG6AN2mkTnm_Wq0Yig)

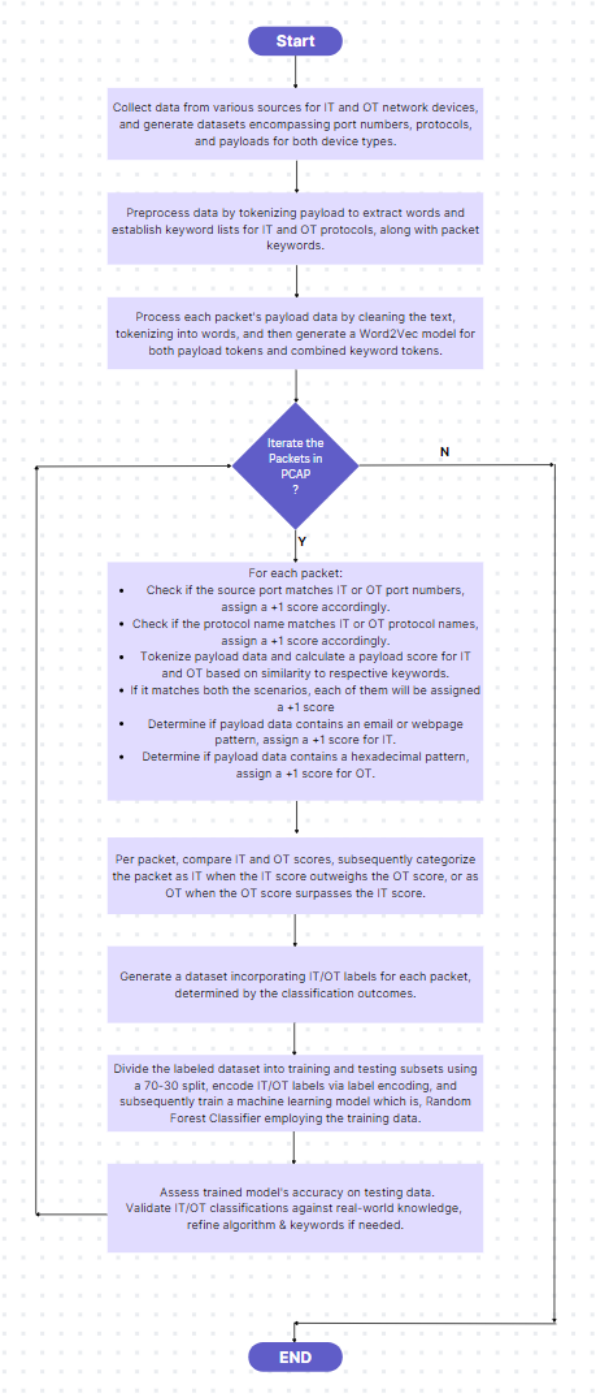
10. Any other paper/literature review pertaining to the same
    - Security Information and Event Management (SIEM): Analysis, Trends, and Usage in Critical Infrastructures.
    - Efficient Passive ICS device discovery and identification by MAC address correlation.
    - On feasibility of Device Fingerprinting in ICS
    - ICS/SCADA Device Recognition: A hybrid communication-patterns and passive fingerprinting approach
    - The problem of security address resolution protocol
    - Many more relevant papers are read and attached below. (drive.google.com/drive/folders/1up7A0DdpeUXHKzJXJrLYIBqc0wZ4gmbM?usp=drive_link)

Challenges and Limitations

Despite the promising approach outlined in this research, several challenges and limitations should be acknowledged. Firstly, the accuracy of the classification heavily relies on the quality and quantity of training data available. The creation of accurate and representative datasets for both IT and OT devices can be time-consuming and require a comprehensive understanding of network protocols and behaviors. Additionally, the proposed scoring system heavily depends on predefined patterns, keywords, and regular expressions. However, the evolving nature of network traffic and the emergence of new communication patterns can lead to false positives or negatives, affecting the classification accuracy. The effectiveness of the Word2Vec-based tokenization and similarity matching may vary based on the diversity and complexity of payload data. Furthermore, the methodology's dependence on external data sources for validation introduces the potential for inaccuracies if the external sources themselves are not completely reliable. Finally, while the Random Forest Classifier offers a practical solution, its performance might be influenced by the feature space, hyperparameter tuning, and dataset size. Addressing these challenges and limitations is vital for enhancing robustness.

Working Flow Model for Classification

In this section, the model seamlessly combines intricate data preprocessing, advanced feature extraction techniques, machine learning algorithms, and content pattern analysis to achieve a granular understanding of network packet characteristics. In this section, we elucidate the key components of the proposed working flow model and its systematic progression, highlighting the meticulous steps involved in realizing accurate and automated IT-OT device classification.



Methodology

1. Data Collection and Preparation

1.1 Data Sources Identification: Identify multiple data sources for both IT and OT network devices. Create datasets containing port numbers, protocol names, and payload details for each dataset.

1.2 Regex Pattern Creation: Develop regular expression patterns for website and email identification in IT devices' payloads, as well as a hexadecimal regex pattern for OT devices.

2. Tokenization and Word2Vec Embedding

2.1 Tokenization of Payload Data: Tokenize payload data for each packet. Convert payload data into lowercase and remove non-alphanumeric characters.

2.2 Word2Vec Model Creation: Utilize the nltk library to tokenize payload data and create Word2Vec models. Model the payload tokens to create numerical embeddings.

2.3 Keywords Incorporation: Combine OT & IT-specific protocol keywords and packet keywords. Train a Word2Vec model on these combined keywords.

3. Scoring System

3.1 Port Number Matching: For each iteration, check if the source port matches the IT or OT port number dataset. If matched, assign a score of +1 to the corresponding network device.

3.2 Protocol Name Matching: Compare the protocol name with IT and OT protocol name datasets. Allocate a score of +1 to the respective network device if a match is found.

3.3 Payload Token Matching: Apply the Word2Vec model on the payload tokens. Determine the similarity between payload tokens and combined keywords. Count the matched tokens and assign a score to IT and OT devices.

4. Payload Content Analysis

4.1 Email and Web Page Detection: Analyze payload data to identify the presence of email or webpage-related patterns. Award a score of +1 to IT devices if an email or webpage pattern is detected.

4.2 Hexadecimal Detection: Examine payload data for the existence of hexadecimal patterns. Provide a score of +1 to OT devices if a hexadecimal pattern is found.

5. Score Aggregation

5.1 Cumulative Scores Calculation: Sum up the scores obtained from port matching, protocol matching, payload token matching, email/webpage detection, and hexadecimal detection for both IT and OT devices.

6. Classification and Model Evaluation

6.1 Data Classification: Classify each packet as an IT or OT device based on the higher cumulative score obtained.
6.2 Cross-Checking Data Integrity: Validate the classification accuracy by cross-referencing the assigned IT/OT label with external sources to ensure data integrity.

7. Data Preprocessing for Machine Learning

7.1 Data Compilation: Collect the calculated scores for each packet to form a dataset.
7.2 Label Encoding: Apply label encoding to convert the categorical IT/OT labels into numerical values suitable for machine learning.

8. Train-Test Data Split and Model Building

8.1 Data Splitting: Divide the dataset into a 70-30 split for training and testing, respectively.
8.2 Random Forest Classifier: Employ the Random Forest Classifier algorithm to train a machine learning model.

9. Model Evaluation

9.1 Accuracy Measurement: Evaluate the trained model's accuracy using the testing dataset.

10. Automation of IT-OT Device Classification

10.1 Automated Classification: Utilize the trained model to automatically classify network devices into IT or OT categories based on their features and scores.

Results and Discussion

The proposed algorithm demonstrated exceptional accuracy, achieving a classification rate of 99.97% in distinguishing between IT and OT network devices. Through a systematic process encompassing feature extraction, scoring, and machine learning, the algorithm effectively harnessed port numbers, protocol names, payload tokens, and content patterns to differentiate between the two categories. The achieved accuracy underlines the algorithm's robustness in automating IT and OT device classification. This high level of accuracy can significantly contribute to enhancing network security and operational efficiency. In conclusion, the algorithm's remarkable accuracy highlights its potential to streamline IT and OT device classification, offering substantial benefits to the cybersecurity landscape.

Future Work

To advance the current research, several promising directions warrant exploration. Firstly, refining the algorithm's adaptability to dynamically evolving network landscapes presents a challenging yet crucial avenue. Incorporating techniques such as online learning and adaptive model updating could bolster the algorithm's resilience to fluctuating network behaviors.

Moreover, extending the algorithm's scope to encompass a broader spectrum of network devices and intricate communication patterns would enhance its applicability in heterogeneous environments. Considering the escalating sophistication of cyber threats, integrating advanced machine learning paradigms like deep learning networks could unravel latent insights within complex packet data, potentially pushing accuracy benchmarks further. To transition the algorithm from research to deployment, a detailed study on real-time implementation, latency considerations, and scalability concerns is imperative. These aspects, coupled with addressing potential privacy issues associated with payload data, would yield a comprehensive framework poised to address intricate network security challenges in diverse operational settings.

Conclusion

In this study, we proposed a systematic methodology for the automated classification of network devices into IT and OT categories, leveraging a combination of feature extraction, scoring mechanisms, and machine learning techniques. By utilizing port numbers, protocol names, and payload data, we developed a scoring system that quantifies the likelihood of a device belonging to either IT or OT networks. The incorporation of Word2Vec-based tokenization enabled the extraction of meaningful features from payload data, enhancing the system's capacity to capture nuanced communication patterns. The obtained results demonstrated the feasibility of the proposed approach, showcasing its potential to effectively differentiate between IT and OT devices with a certain degree of accuracy. However, we acknowledge the challenges posed by the dynamic nature of network traffic and the need for high-quality training data. Further improvements could involve refining the scoring mechanism, exploring more advanced machine learning algorithms, and devising strategies to accommodate evolving communication patterns. As the landscape of network devices continues to evolve, the findings of this research contribute to the ongoing discourse on automated device classification and pave the way for enhanced security, management, and optimization of heterogeneous networks.