Opanovych Maksym, PhD student

National University ‘Lviv Polytechnic’, Lviv

Zhuravchak Danylo, PhD student

National University ‘Lviv Polytechnic’, Lviv

Dudykevych Valerii, Doctor of Technical Sciences,

Professor, Head of the Department

Lviv Polytechnic National University, Lviv

Andrian Piskozub, Associate Professor

Lviv Polytechnic National University, Lviv

Data mining for threat detection in Active Directory

**Ключові слова: моніторинг безпеки видобування даних, Active Directory**

**Keywords: security monitoring, data mining, Active Directory**

**Introduction**

Cyber threats and malware have evolved significantly over the past few decades. Initially, malware was relatively simple, often created by individuals to experiment or demonstrate technical prowess. Early viruses and worms spread through floppy disks and basic network connections. However, with the growth of the internet, cyber threats have become more sophisticated and widespread. Today, organized crime groups and state-sponsored actors often develop malware, targeting individuals, corporations, and governments. Advanced techniques such as polymorphic malware, ransomware, and zero-day exploits have emerged, making detection and prevention increasingly challenging. As cyber threats continue to evolve, so too must the strategies and technologies used to combat them [1].

Over time, many organizations began to recognize the value of collecting and analyzing vast amounts of data from their infrastructure. Initially gathered for operational insights and efficiency improvements, this big data eventually found critical applications in the realm of cybersecurity.[2] As the volume and complexity of cyber threats grew, the ability to process and analyze extensive datasets became essential for detecting and mitigating security incidents.

Data mining has become an essential technique for detecting malicious activity by leveraging the vast amounts of data collected from various sources within an organization's infrastructure.[3] This process involves analyzing large datasets to uncover patterns, correlations, and anomalies that may indicate security threats. By applying data mining algorithms, organizations can sift through massive volumes of logs, network traffic, and user activities to identify suspicious behavior that traditional security measures might miss.[4] The integration of big data enhances the accuracy and speed of these detections, allowing for more proactive and effective cybersecurity strategies.[5]

Tools like Splunk and IBM’s QRadar are designed to harness the power of data mining for cybersecurity purposes. Splunk, for instance, uses its advanced analytics capabilities to parse and analyze log data in real-time, enabling it to detect unusual patterns that could signify malware infections, data breaches, or Advanced Persistent Threat (APT) activities. By correlating events across multiple datasets, Splunk can identify complex attack vectors that would be difficult to spot through manual analysis. This real-time analysis is crucial for organizations to respond quickly to potential threats and minimize the impact of security incidents.

IBM’s QRadar also exemplifies the use of data mining in detecting malicious activity. By collecting data from diverse sources such as network traffic, user activities, and application logs, QRadar applies machine learning and AI algorithms to identify deviations from normal behavior. This capability allows QRadar to detect a wide range of threats, from simple malware infections to sophisticated APT campaigns. The platform’s ability to analyze and correlate large datasets in real-time enhances its effectiveness in identifying and mitigating security risks. Through these advanced data mining techniques, QRadar provides organizations with a robust toolset for protecting their infrastructure against an ever-evolving threat landscape.

**Data mining for threat detection in Active Directory**

Data mining can be a powerful tool for detecting threats within Active Directory (AD) environments by analyzing the vast amount of data generated by user activities, access logs, and system events. Active Directory is a critical component of many organizational IT infrastructures, managing user authentication, authorization, and directory services. Due to its central role, AD is often a target for cyber attacks, making its security paramount.

By applying data mining techniques, organizations can sift through the extensive logs and event data generated by Active Directory to identify patterns and anomalies indicative of malicious activity. For instance, unusual patterns in login attempts, sudden changes in user permissions, or the creation of unauthorized accounts can be flagged as potential security threats. Data mining algorithms can analyze these patterns across various dimensions, such as time, user roles, and access levels, to detect deviations from typical behavior.

Data mining is invaluable for threat detection due to its ability to process and analyze vast amounts of data to uncover hidden patterns, correlations, and anomalies that traditional methods might miss. As cyber threats become increasingly sophisticated and diverse, relying solely on conventional security measures is no longer sufficient. Data mining leverages advanced algorithms and machine learning techniques to identify subtle indicators of potential threats, providing a deeper and more comprehensive understanding of security events.

One of the primary reasons data mining is crucial for threat detection is its capability to handle the enormous volume and variety of data generated by modern IT infrastructures. Every user action, network transaction, and system event produces data that, when analyzed collectively, can reveal insights into potential security threats. Data mining enables organizations to sift through this massive data efficiently, identifying patterns that suggest malicious activity. This is particularly important in detecting complex and multi-faceted attacks, such as Advanced Persistent Threats (APTs), which may involve coordinated actions across different parts of the network.

Furthermore, data mining enhances the accuracy and speed of threat detection. By applying machine learning and artificial intelligence, data mining tools can continuously learn from new data, improving their ability to recognize emerging threats. This adaptive learning process allows for real-time analysis and immediate detection of anomalies, significantly reducing the time between the occurrence of a suspicious event and its identification. Consequently, organizations can respond more swiftly to mitigate potential damage, enhancing their overall cybersecurity posture. In summary, data mining is a powerful tool for threat detection, providing the scalability, speed, and precision needed to safeguard against evolving cyber threats.



Figure 1: Data mining Flowchart

Data mining involves a systematic series of steps designed to extract valuable insights from large datasets from different data sources as it represented in Figure 1. The process begins with data collection and integration. Organizations gather data from various sources such as databases, log files, sensors, and other repositories. This collected data is then integrated into a unified dataset, often requiring the merging of data from multiple databases or the integration of data from different formats. This step ensures that all relevant data is available in a cohesive form, and ready for analysis.

The next step is data cleaning, which involves preprocessing the data to identify and correct errors, remove duplicates, and fill in missing values. This step is crucial to ensure the accuracy and completeness of the data, as any inaccuracies can lead to erroneous insights. Additionally, noise removal is performed to filter out irrelevant or redundant information that may obscure important patterns. This refined data is then transformed through normalization, which scales the data to a standard range to ensure consistency in analysis. Aggregation may also be performed to summarize the data into higher-level forms, such as computing averages or sums, while discretization converts continuous data into discrete intervals for easier analysis.

Following data transformation, the data is reduced through techniques such as feature selection and data compression. Feature selection identifies and retains the most relevant attributes for analysis, reducing the dimensionality of the dataset. Data compression techniques like sampling or clustering further reduce the size of the dataset, making processing more efficient. The core of data mining involves model building, where data mining algorithms are applied to construct models that can identify patterns and relationships within the data. These patterns are then evaluated to determine their significance and relevance.

Model evaluation and validation are critical to ensure the accuracy and performance of the models. This is achieved by testing the models using a separate test dataset and employing cross-validation techniques, such as k-fold cross-validation, to ensure the model's robustness and generalizability. Once validated, the discovered patterns and insights are represented through visualization tools like charts, graphs, and dashboards, making them accessible and comprehensible. Comprehensive reports are generated to summarize the findings and provide actionable insights.

Finally, the deployment phase involves integrating the validated models into operational systems for real-time monitoring and decision-making. Continuous monitoring of the model’s performance is essential, with updates made as necessary to adapt to new data and evolving threats. This cyclical process, from data collection to deployment, allows organizations to systematically extract meaningful patterns and insights from their data, significantly enhancing their ability to detect and respond to various threats.

In Active Directory environment data mining can be effectively utilized for malware and anomaly detection by analyzing logs from Windows, Active Directory (AD), and Sysmon (System Monitor). These logs provide a wealth of information about system activities, user behaviors, and network traffic, which can be mined to identify suspicious patterns and potential security threats.

Windows logs, such as event logs, contain detailed records of system, security, and application events. Data mining techniques can be applied to these logs to detect anomalies that may indicate malware activity.

For instance, unusual patterns in login attempts, such as repeated failed logins followed by a successful login, can indicate a brute force attack. Another example is monitoring for the execution of rarely used system utilities (e.g., regsvr32, powershell, wmic) in unusual contexts, which can be indicative of malware exploitation. Clustering algorithms can group similar events and highlight outliers, such as a spike in the number of process terminations or the sudden appearance of a new process that begins executing suspicious activities.

Sequence mining can track the order of events to detect patterns commonly associated with malware. For example, detecting a sequence where a user logs in, followed by the execution of a command that alters the Windows registry, and then by a sudden network connection to an external IP address, can be indicative of a malware infection trying to establish a command and control channel.

Active Directory is a critical component for managing user access and permissions in a Windows environment. Analyzing AD logs can help detect unauthorized access attempts, privilege escalation, and lateral movement within the network.

For example, monitoring login attempts can reveal anomalies such as logins from unusual locations or at unusual times, suggesting potential compromised accounts. Another instance is detecting changes to group memberships, such as a non-administrative user being added to a privileged group, which can indicate privilege escalation attacks. Analyzing account lockout events can also reveal potential brute force attacks on user accounts.

Association rule mining can uncover relationships between seemingly unrelated events in AD logs. For example, a new user account being created and immediately assigned administrative privileges can be suspicious. Similarly, frequent password reset requests for privileged accounts can indicate an attacker attempting to gain control over critical accounts.

Sysmon provides detailed logging of system activity, including process creation, network connections, and changes to file creation timestamps. Data mining on Sysmon logs can reveal intricate details about potential malware operations.

For example, anomaly detection algorithms can flag rare or unexpected process executions, especially those involving known malicious executables or scripts. Monitoring for the creation of processes from unusual paths, such as executables running from the AppData or Temp directories, can indicate malware activity. Detecting changes in file creation timestamps that do not match the typical usage patterns can suggest file tampering by malware.

Machine learning models can be trained to recognize the signatures of known malware behaviors. For instance, identifying processes that initiate a large number of network connections in a short time frame can be indicative of a botnet or worm. Monitoring for the usage of tools like Procmon or Process Explorer in contexts where they are not typically used can also reveal sophisticated attacks trying to avoid detection.

Combining data from Windows, AD, and Sysmon logs provides a comprehensive view of the security landscape. Data mining techniques can correlate events across these different log sources to identify complex attack patterns that might be missed when analyzing logs in isolation.

For example, an unusual network connection detected in Sysmon logs can be correlated with a suspicious user login in AD logs and an unexpected process execution in Windows logs. This integrative approach can reveal a multi-stage attack where an attacker gains initial access through a phishing email (Windows log), escalates privileges using AD (AD log), and then moves laterally within the network while exfiltrating data (Sysmon log).

In another scenario, a sequence of events might involve a non-privileged user suddenly accessing sensitive files (Windows logs), followed by changes to group memberships (AD logs), and ending with the initiation of a remote desktop session to an external IP (Sysmon logs). By correlating these events, organizations can detect advanced persistent threats (APTs) that use sophisticated techniques to evade traditional detection methods.

**References**

1. Gupta, S., Sabitha, A.S., Punhani, R. Cyber security threat intelligence using data mining techniques and artificial intelligence. Int. J. Recent Technol. Eng, 2019. [Electronic resource]. Available at: <https://www.ijrte.org/portfolio-item/C5675098319/>
2. Foroughi, F., Luksch, P. Data science methodology for cybersecurity projects. arXiv preprint arXiv:1803.04219, 2018. [Electronic resource]. Available at: <https://arxiv.org/pdf/1803.04219>.
3. Salem, I.E., Mijwil, M.M., Abdulqader, A.W. Introduction to the data mining techniques in cybersecurity. International Journal of Cybersecurity, 2022. [Electronic resource]. Available at: <https://www.iasj.net/iasj/download/ad6291d5f5f3cd24>.
4. Dua, S., Du, X. Data mining and machine learning in cybersecurity. Books.google.com, 2016. [Electronic resource]. Available at: <https://books.google.com/books?hl=uk&lr=&id=1-FY-U30lUYC&oi=fnd&pg=PP1&dq=cybersecurity+data+mining&ots=cHFjH9Qgf3&sig=faQ-7W0LjaT87c__JKHUtWb6CpE>.
5. Buczak, A.L., Guven, E. A survey of data mining and machine learning methods for cyber security intrusion detection. IEEE Communications surveys & tutorials, 2015. [Electronic resource]. Available at: <https://ieeexplore.ieee.org/abstract/document/7307098/>.