

Thèse qui présente des dataset : Lateral Movement Attacks Datasets: Benchmarking, Challenges, and Solutions (6 décembre 2024)

Objectif de ce texte :

However, the detection of lateral movement has been hindered by a lack of comprehensive, high-quality datasets that accurately reflect the diverse and evolving tactics used in such attacks. Existing datasets suffer from several limitations, including a scarcity of lateral movement instances, outdated attack patterns, and insufficient diversity in techniques and attack paths, especially in cloud-based environments. Moreover, automatic labeling methods for dataset creation are often imprecise, complicating the training of effective detection models. This work addresses these challenges by proposing a new benchmark dataset specifically tailored for lateral movement attacks. We conduct a comprehensive analysis of existing lateral movement attack datasets, highlighting gaps and providing insights into the strengths and weaknesses of current approaches. In response, we introduce the Lateral Movement Dataset Generator (LMDG), a framework designed to generate high-quality datasets for lateral movement and APT detection.

Pourquoi faire ? :

Current research endeavors for lateral movement detection rely on machine learning [74, 12, 53, 77, 55]. The machine learning paradigm depends heavily on datasets to train and evaluate detection models, and the quality of these datasets directly impacts model performance and evaluation accuracy. Without high-quality training data, models can exhibit performance discrepancies, reducing accuracy and increasing false positives [25, 45] (see section 3.2).

Définition de la latéralisation :

Lateral movement, a pivotal tactic within APT campaigns, allows attackers to pivot across networked systems, escalating privileges and accessing critical resources that were initially out of reach. As cyber adversaries continue to refine their techniques, lateral movement has emerged as a key strategy for evading detection, maintaining access, and progressing toward their ultimate goals [1, 9, 19, 17].

Défis de la détection de la latéralisation :

The challenge in detecting lateral movement is compounded by several factors. These attacks are often prolonged, with threat actors moving through networks over extended periods while blending in with normal network activity. Additionally, the sheer volume of data generated by enterprise networks makes it difficult to identify malicious activity amidst the noise. Attackers frequently exploit legitimate authentication credentials and system tools, further obscuring their actions. The complexity of detecting lateral movement is also heightened by the use of novel malware variants, zero-day exploits, and evasion techniques that allow attackers to bypass conventional detection mechanisms [5, 11, 7, 4, 3]. The growing sophistication of lateral movement tactics has made it a critical focus for cybersecurity research, with numerous efforts aimed at developing models for its early detection and mitigation [6]. Despite the importance of lateral movement detection, current research is hindered by challenges related to data quality. The effectiveness of machine learning (ML) models for detecting lateral movement depends heavily on the quality and accuracy of the datasets used for training and evaluation. Many existing datasets suffer from issues such as noisy labels, class imbalances, and insufficient diversity in attack patterns, limiting their usefulness for developing robust detection models [8, 13]. Furthermore, most datasets lack sufficient instances of lateral movement attacks, making it difficult to train models that can generalize across a wide range of attack scenarios [20, 10].

Importance d'une détection rapide :

Since lateral movement is a crucial phase in an APT attack, early detection is vital to minimize losses and prevent attackers from gaining further access to the network [10].

Progression horizontale vs verticale :

Let's begin by establishing two fundamental concepts: horizontal progression and vertical progression. Horizontal progression entails obtaining an initial foothold on one or multiple hosts within a network, with each initial access executed independently of the others. For instance, consider a scenario where a network scan reveals ten hosts within a segment. Among these, three have distinct vulnerabilities that can be exploited for access. It's important to note that each of these initial accesses occurs in isolation. This form of horizontal progression, while significant for gaining initial access, does not qualify as lateral movement. On the contrary, vertical progression involves accessing multiple systems where these accesses are interdependent. To illustrate, an adversary might secure an initial foothold within network segment A, proceed to segment B, and then advance to segment C. Importantly, these initial accesses are not isolated but instead rely on one another. For instance, the adversary gains control over a host in segment A, providing remote access to a machine in segment B. From a machine in segment B, further access is obtained to a machine in segment C. Lateral movement, therefore, can be defined as the vertical progression between hosts, accounts, or the transition from one set of privileges to another.

Mouvement latéral dans l'infonuage :

One key distinction is that, instead of having hosts or resources directly, there exists an additional layer known as the services layer. These services, like AWS EC2 or S3 buckets, offer resources such as compute instances and object storage. Consequently, in the context of lateral movement in the cloud, we observe a vertical progression encompassing identities, permissions/policies, services, and resources. A recent example highlighted by the Microsoft Threat Intelligence team [6] involved adversaries gaining initial access to an Azure-based database server through SQL injection. Subsequently, they attempted to obtain a cloud identity token using the IMDS (Instance Metadata Service) to access other cloud resources.

Étude des jeux de données :

- **LANL Datasets (2015, 2018) :** Pas de mouvement latéral
- **DARPA Transparent Computing Engagement 3 (DARPA 2018) :** Pas de mouvement latéral
- **DARPA Transparent Computing Engagement 5 (DARPA 2019) :** 2 scénarii proches de mouvement latéral
- **DARPA Operationally Transparent Cyber 2019 (OpTC) :** 2 scénarii de mouvement latéral avec plus de transitions que dans le cas précédent.
- **CERT Insider Threat Dataset (2020) :** *comprehensive collection of real-world instances and data related to insider threats*, pas de mouvement latéral
- **PicoDomain Dataset (2020) :** *Notably, lateral movement (LM) predominantly utilized WMI and DCOM techniques with minimal diversity. The dataset comprises a single LM scenario, shown in figure 2.2.2, over a short three-day span, lacking instances with an extended temporal scope.*
- **DARPA Intrusion Detection Datasets (1998, 1999, 2000) :** Pas de mouvement latéral
- **NDSec-1 Dataset (2017) :** *It's noticeable that this dataset comprises just two scenarios, both executed within a single day. The lateral movement path's extent can be seen as encompassing two hops.*
- **Pivoting Detection Dataset (2017) :** *Upon creating a graph that visualizes all the pivoting activities with their temporal progression within the dataset in figures 2.2.3 2.2.4 2.2.5 2.2.6, it becomes evident that there are only a few instances of Lateral Movement, all of which involve two hops. The specific technique employed in these pivoting instances remains unclear; however, it is noteworthy that all of them were executed over TCP.*
- **StreamSpot Dataset (2016) :** Pas de mouvement latéral
- **ISCX Intrusion Detection Evaluation Dataset (2012) :** *The paper's description of the attack*

scenarios reveals four distinct attacks spanning a week. Upon analyzing these attacks, it becomes apparent that each case involved a series of lateral movements, encompassing two successive hops, passing through three hosts. The sequence of lateral movements consistently began with the targeted hosts receiving a deceptive email containing a malicious PDF file harboring a reverse TCP shell. Subsequently, both the initial and secondary hops followed a uniform approach, exploiting hosts operating Windows XP and utilizing a vulnerable SMB authentication protocol. Notably, there was only a single occurrence where the second hop was executed through a brute force method to ascertain user credentials.

- **DAPT (2020)** : However, the dataset includes only a lateral movement instance executed over a 10-hour window—a significantly shorter timeframe than realistic LM attacks, which can span weeks or even months.
- **Unraveled (2023)** : However, the lateral movement phase remained relatively simplistic, occurring within a single day and consisting of internal reconnaissance and password cracking. In both the DAPT2020 and Unraveled datasets, the attack execution and labeling processes were conducted manually.

In summary, the current datasets face significant challenges, highlighting the need to develop a new dataset that effectively addresses these issues. These challenges encompass a shortage of Lateral Movement instances in existing datasets, which hinders model training and generalization. Furthermore, the limited diversity in Lateral Movement techniques and the often short timeframes associated with these activities present additional obstacles in creating robust detection models. Moreover, the prevalence of Lateral Movement paths comprising only a few hops limits the current datasets' scope and symbolic value. The absence of dedicated datasets tailored for Cloud-based Lateral Movement scenarios further underscores the need for comprehensive and up-to-date resources in this domain. Another critical concern is the obsolescence of existing datasets, rendering them inadequate for capturing recent attack patterns and trends. Lastly, some datasets offer only a partial view of the overall threat landscape, focusing solely on network flow data, thus emphasizing the necessity for more comprehensive datasets encompassing a broader spectrum of Lateral Movement activities.

LMDG: A Framework for Lateral Movement Datasets Generation :

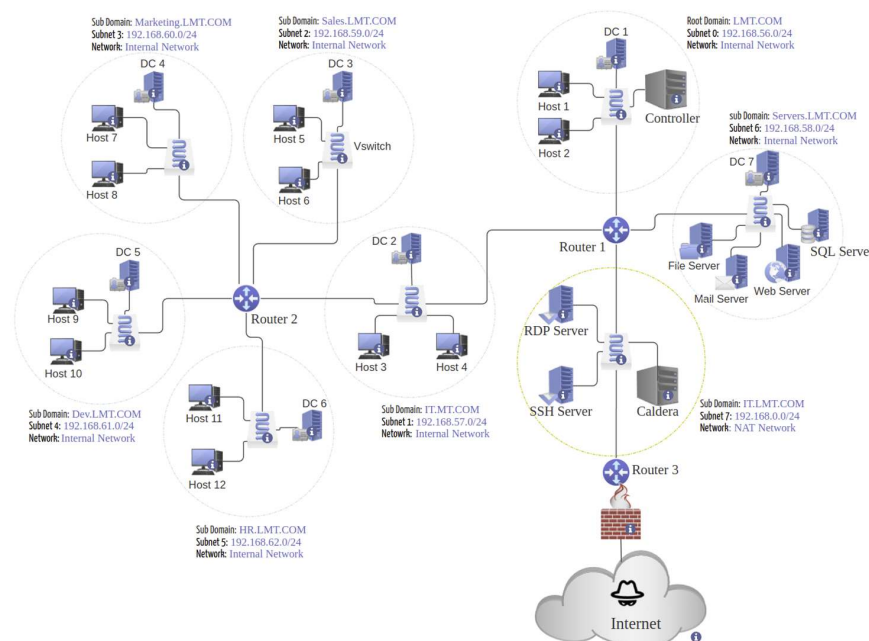


Fig. 3.3.1: The network topology used to generate LMDG dataset.

Following the approach outlined in [33], we utilized Caldera to implement the client server architecture discussed in subsection 3.3.4.3 to automate attack execution steps. Ideally, the Caldera Server should

operate without leaving any trace of its communication with the agents or other artifacts that could influence the dataset. The server should remain transparent and invisible to avoid contaminating the dataset with automation-related traces. Attack scenarios that involve an actual C2 server should be hosted on a separate machine from the Caldera Server, and all Caldera Server artifacts should be removed from the dataset to maintain its integrity.

The LMDG dataset contains seven attack scenarios that achieve lateral movement using various tactics and techniques. Of these, three attacks were successful, while four were unsuccessful. We discuss possible reasons for each unsuccessful attack, considering that our setup includes Windows 10, Windows 11, and Windows Server 2022—the latest Windows versions with advanced security mechanisms. This combination of successful and unsuccessful attacks is valuable for understanding attacker behavior, as many attacks tend to fail due to robust defenses, with only some achieving success.

The total compressed dataset size, encompassing benign and malicious data (excluding router data), is 253 GB; when router data is included, the dataset size increases to 527 GB. Specifically, the compressed PCAP file from router 1 is 201 GB, and that from router 2 is 72 GB. The total uncompressed dataset amounts to 944 GB, with 900.93 GB comprising PCAP files and 43.38 GB for system log files. Additional dataset statistics for the uncompressed data are presented in Table 3.4.1.

=> LMD-2022 missing