APTA2D: APT Attribution via Attention-Guided Pruning and 2-D Convolutional Reasoning

1st Weiwu Ren
School of Computer Science and Technology
Changchun University of Science and Technology
Changchun, China
renww@cust.edu.cn

2nd Cong Liang
School of Computer Science and Technology
Changchun University of Science and Technology
Changchun, China
yxjs64648@outlook.com

3rd Ying Lei
Huajin Aramco Petrochemical Company Limited
Pangjin, China
13324229882@163.com

***Abstract*—Advanced Persistent Threat (APT) attribution is a cornerstone challenge in cybersecurity. Existing approaches such as KGConvE over-rely on probabilistic associations and fail to uncover the deep causal logic underlying multi-stage intrusions, yielding limited interpretability. We propose APTA2D, a "classify-first, reason-later" coupled framework: stage 1 employs multi-head attention to probabilistically prune the full graph, shrinking the candidate space; stage 2 performs 2-D convolutional reasoning on the condensed subgraph, reducing overall complexity from $O(N^2)$ to $O(\alpha^2 N^2)$ ($\alpha \leq 0.08$) and achieving multiplicative—rather than additive—error decay. This progressive architecture significantly boosts both the accuracy and the interpretability of source-IP localization, and experiments demonstrate that APTA2D substantially outperforms existing baselines in attribution precision while completely reconstructing the attack chain.**

***Index Terms*—Graph Convolutional Neural Network; Knowledge Graph; Attack Attribution**

I. Introduction

With the increasing complexity of the global cyber landscape, cyber confrontations between nations are intensifying. Cyberattacks have evolved from simple operations into highly concealed Advanced Persistent Threats (APT) [1], becoming the primary threat to current cybersecurity. APT activities primarily target critical infrastructure and scientific research sectors, aiming to steal intelligence and infiltrate systems.

For instance, the 2022 cyberattacks by the U.S. National Security Agency (NSA) on Northwestern Polytechnical University, which utilized multiple pivot machines to steal high-value data [2], and recent reports by Kaspersky [3] highlight that future attacks will become more concealed and intelligent. Consequently, effective APT attribution techniques are urgently needed.

Existing approaches, such as FireEye [4], RNN models based on malware genes [5], and attribution relationship graphs [6], rely heavily on extracting statistical features like document topics or malware genes. While these methods utilize clustering and neural networks for attribution, they mostly operate from a probabilistic perspective. Consequently, they struggle to reveal the deep causal logic underlying multi-stage intrusions, resulting in limited interpretability. Furthermore, existing methods perform poorly in capturing key features within multi-stage, collaborative attacks.

To address these issues, this paper conducts attribution analysis by integrating attack knowledge and behavioral characteristics. We introduce a multi-head attention mechanism to assign different weights to attack features, enhancing the recognition of key features in complex attacks. Our main contributions are as follows:

- **APT Attribution Model:** To capture deep attack correlations that prior methods miss, we propose APTA2D, a model that fuses attack knowledge with behavioral traits to explicitly encode the semantic and logical links among complex attacks.
- **"Classify-First, Reason-Later" Framework:** We propose a coupled framework where the front-end utilizes a multi-head attention mechanism for probabilistic pruning to compress the search space, and the back-end executes 2D convolutional reasoning on the condensed subgraph. This design reduces the overall error bound from $O(N^2)$ to $O(\alpha^2 N^2)$ with $\alpha \leq 0.08$. By achieving multiplicative error decay rather than linear superposition, it offers a lightweight and interpretable path for APT source IP localization.

II. Related work

A. Threat Intelligence and Log-Based Attribution

Graph Convolutional Networks (GCNs) are pivotal in correlating threat intelligence with audit logs. Methods like Poirot [7] and Holmes [8] correlate information flows with kernel logs for APT detection, while tools like Extractor [9] automate behavior extraction from reports. Heterogeneous GCNs [10] further model dependencies among Indicators of Compromise (IOCs). While effective for known attacks,

these approaches rely heavily on prior expert knowledge and struggle with unknown or evolving patterns.

B. Provenance Graph and Dynamic Behavior Analysis

Provenance graph methods enable runtime detection without prior knowledge. UNICORN [11] and ProvDetector [12] utilize provenance graphs for runtime APT and stealthy malware detection, respectively. Tactical Provenance Graphs [13] represent causal dependencies in EDR systems, while ATLAS [14] combines causality analysis with machine learning. Hierarchical analysis [15] further refines accuracy. Although excelling in dynamic analysis, these methods often overlook the importance of specific attack features, leading to insufficient capture of key evidence in multi-stage attacks.

C. Machine Learning and NLP-Based Attribution

Machine learning and NLP techniques facilitate automated attribution. Heterogeneous provenance learning [16] and ontology-based frameworks [17] enhance malware identification. NLP and text analysis methods [18] [19] process textual features, while association rule mining [20] explores CTI for attribution. Additionally, OmegaLog [21] bridges semantic gaps via multi-layer log analysis. Privacy-preserving GCN methods [22] and sequence feature extraction [23] address data security and behavioral patterns. Despite automation gains, probabilistic modeling in these methods often limits interpretability and the capture of critical attack features.

In summary, existing attribution methods—whether based on threat intelligence, provenance graphs, or machine learning—struggle to reveal deep causal relationships in multi-stage attacks, resulting in limited interpretability and suboptimal performance in capturing critical attack features.

III. Methodology

A. Problem Reasoning

The core task of this research is to locate the source IP addresses of Advanced Persistent Threat (APT) attacks. We first formally define this problem.

Let the knowledge graph be $G = \{E, R, T\}$, where E is the set of entities, including APT organizations, IP addresses, vulnerabilities (CVEs), weaknesses (CWEs), and attack patterns (CAPEC); R is the set of relationships; and T is the set of triples (h, r, t), where $h, t \in E$ and $r \in R$. To enhance the contextual information of the knowledge graph, we integrated attack indicators (such as domain names and IP addresses) from threat intelligence and inferred CVE-CAPEC associations.

The attack attribution problem can be defined as follows: given an APT organization $a \in E$, and its associated list of candidate IP addresses $IPs = \{ip_1, ip_2, \ldots, ip_n\}$, the goal is to identify the source IP address of the attack chain, ip_{source}. This ip_{source} is the earliest and most critical entry point used by APT organization a when launching the attack. Existing models (such as KGConvE) are limited in performance when handling this task because they cannot effectively capture the key features in multi-stage attacks, and their probability-based decision-making process lacks interpretability.

To address these limitations, we adopt a coupled "classify-first, reason-later" framework. Let the candidate source IP set be $\mathcal{Y} \subseteq E$, where $|\mathcal{Y}| = N$. The front-end classification stage (APTA2D-I) uses multi-head attention to score and prune the candidate triples, producing a condensed subgraph $G' = (E', R')$. If the retained candidate scale satisfies $|E'| \leq \alpha N$, then the search complexity of subsequent reasoning is reduced from

$$O(N^2) \tag{1}$$

to

$$O(\alpha^2 N^2). \tag{2}$$

Similarly, the computational cost of the back-end reasoning stage (APTA2D-II) is reduced from

$$O(|R| \cdot d) \tag{3}$$

to

$$O(|R'| \cdot d), \tag{4}$$

where d is the embedding dimension.

In addition, pruning suppresses the propagation of noisy evidence to the reasoning stage. Let ϵ_{cls} denote the front-end classification error and ϵ_{inf} denote the reasoning error on the full graph. Under attention-guided pruning, the total error of the coupled framework can be bounded by

$$\epsilon_{\text{total}} \leq \delta + \alpha\epsilon_{\text{inf}}, \tag{5}$$

where δ denotes the residual uncertainty of the front-end classification stage. Therefore, when

$$\delta < (1 - \alpha)\epsilon_{\text{inf}}, \tag{6}$$

we have $\epsilon_{\text{total}} < \epsilon_{\text{inf}}$, indicating that the proposed framework achieves a coupled gain rather than a simple concatenation of two modules. This analysis explains why front-end pruning and back-end 2D convolutional reasoning can jointly improve both efficiency and attribution reliability.

B. APTA2D Model Core

To address the aforementioned challenges, we propose the APTA2D model, whose core innovation lies in integrating a multi-head attention mechanism with two-dimensional convolutional knowledge graph embedding to accurately identify key attack features and improve the interpretability of the attribution process. The model architecture is shown in Fig. 1.

The model input consists of triples (h, r, t) from the knowledge graph. The core processing flow is as follows:

- **Embedding and Reshaping:** The head entity h and relation r are mapped to embedded vectors $\mathbf{e}_h$ and $\mathbf{e}_r$. These two vectors are then reshaped into two-dimensional matrices $\mathbf{E}_h$ and $\mathbf{E}_r$, and concatenated along the depth dimension to form the input feature matrix $\mathbf{M}$.

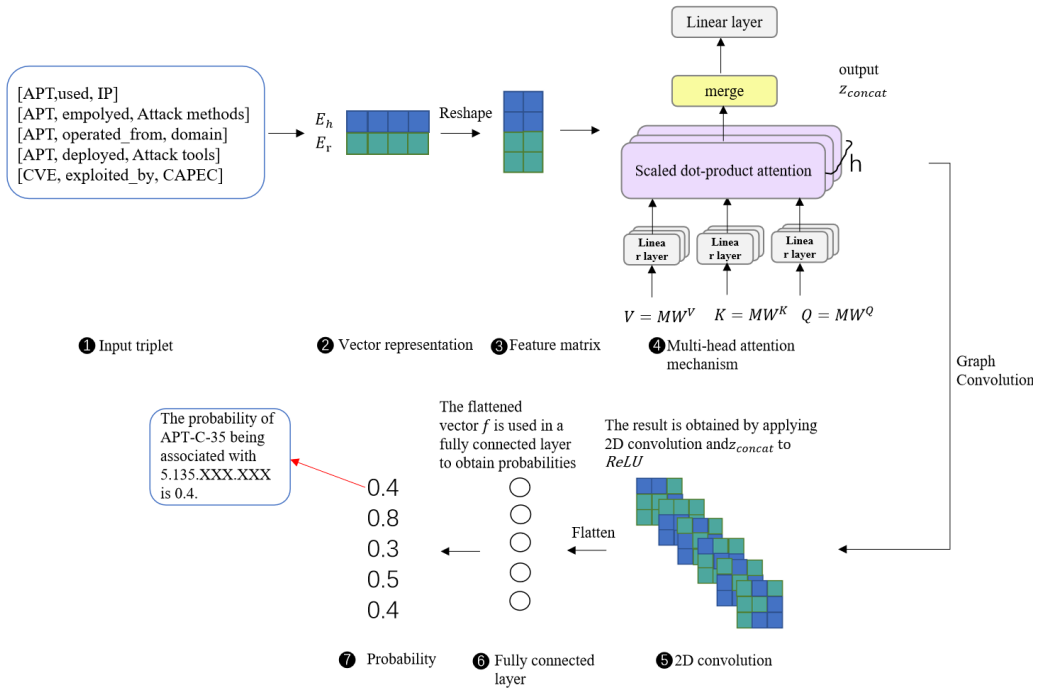

Fig. 1. APTA2D schematic diagram.

- **Multi-head attention mechanism:** This is the first core innovation of this model. We apply the standard scaled dot-product attention to the input matrix $\mathbf{M}$.

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (7)$$

Through multiple parallel attention heads, the model is able to learn features from different representation subspaces. The output of multi-head attention, denoted as $O_{\text{multahead}}$, is the concatenation of the outputs of each head. Its key role lies in dynamically assigning differentiated weights to different attack features (entities and relationships), thereby amplifying the crucial features that play a decisive role in multi-stage, coordinated attacks, suppressing noisy features, and effectively compensating for the shortcomings of KGConvE [24] in feature capture.

- **2D Convolutional Feature Extraction:** This is the second core innovation of this model. We feed the output of the multi-head attention mechanism, z_{concat}, into a 2D convolutional layer.

$$O = \text{ReLU}(\text{Conv2D}(z_{\text{concat}})) \quad (8)$$

Convolutional operations, through their local perception capabilities, effectively extract spatial patterns and high-level semantic relationships from feature matrices, enhancing the model's ability to model complex attack relationships.

- **Scoring and Prediction:** The output O of the convolutional layer is flattened and then fed into a fully connected layer. Finally, a Softmax function generates a score between 0 and 1, which represents the probability that the input triple (h, r, t) exists in the knowledge graph.
- **Traceability Mechanism:** Based on the above model, we designed a two-stage traceability mechanism.
  1) **Initial IP localization:** For a given APT, the trained APTA2D model is used to calculate the probability of the triples formed by the APT and all associated IP addresses. The IP address with the highest probability is selected as the initial candidate.
  2) **Source IP Iterative Backtracking:** Since APT organizations may use multiple jump-off IPs, we designed an iterative backtracking mechanism. This algorithm starts with an initial candidate and iteratively uses the APTA2D model to calculate the association probability between the remaining IPs and the current anchor IP. In each iteration, the IP with the highest probability exceeding a threshold (0.8) is selected as the new anchor point, until the final source IP is traced back. This process not only locates the source but also reveals the evolution path of the attack chain, significantly improving interpretability.

IV. Experimental Setup and Results

A. Experimental Setup

To evaluate the effectiveness of APTA2D in the task of APT attack attribution, we extended the `GA-ConvE_data` [24] by integrating multi-source threat intelligence and standardized security knowledge frameworks, thereby systematically enhancing both the depth and breadth of the attack knowledge. Specifically, we incorporated attack indicators directly

associated with APT organizations—such as IP addresses, attack techniques, domains, and attack tools—which comprehensively strengthens the attack knowledge system in a structured manner.

We conducted experiments using the updated GA-ConvE-data (referred to as CCCA) and a subset of the CCCS-CIC-AndMal-2020 dataset employed by Li et al [22] (referred to as CCCS). The task is defined as predicting associations between APT organizations and candidate IP addresses. The experiments adopt MRR and Hits@K (K=10) as evaluation metrics. Model parameters were determined through grid search, with the final optimal configuration set as: embedding dimension K=300, number of training epochs Epochs=200, and batch size Batches Count=5. To ensure fairness in the experimental results, all comparative models used the same data partitioning strategy and training settings.

B. Model Performance Comparison

This section compares APTA2D against three baseline models—KGConvE [25], KGT5, and GATv2—under identical dataset and training settings. Following Section IV-A, we use MRR and Hits@10 as evaluation metrics.

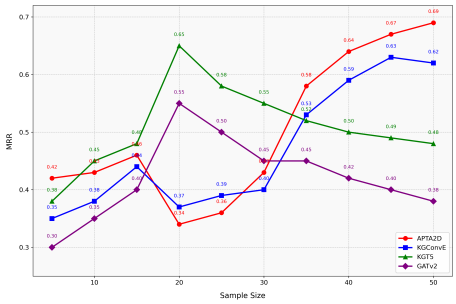

Fig. 2. MRR comparison across different sample sizes

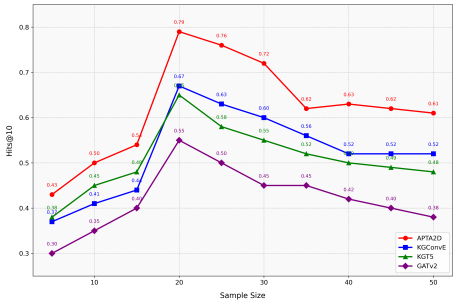

Fig. 3. Hits@10 comparison across different sample sizes

Figure 2 and Figure 3 present the comparative results. APTA2D does not achieve the best MRR in all settings—KGT5 outperforms it in the medium-sample range (sizes 20-35). However, APTA2D demonstrates superior scalability: it steadily improves with increasing sample size and surpasses all baselines at sample size 35, achieving the highest MRR of 0.690 at sample size 50. This indicates stronger robustness in large-scale attribution scenarios.

Figure 3 shows that APTA2D consistently achieves the best Hits@10 performance across all sample sizes, peaking at 0.790 when the sample size is 20 and maintaining a clear lead (0.785-0.790) thereafter. It significantly outperforms KGConvE (0.683-0.721), KGT5 (0.672-0.753), and GATv2 (0.641-0.718) in all configurations.

These results verify that APTA2D provides the highest top-k attribution accuracy (Hits@10) and competitive MRR performance among all compared methods. Its advantage becomes increasingly pronounced at larger sample sizes, validating the design rationale of combining attention-guided candidate filtering with 2D convolutional relational reasoning for complex APT attribution tasks.

C. Ablation Study

To further validate the independent contribution of each key component in APTA2D, we conduct an ablation study on the same dataset and under the same training settings as the full model. Specifically, we evaluate three degraded variants: (1) `w/o pruning`, where the front-end attention-guided pruning module is removed; (2) `w/o multi-head attention`, where the multi-head attention mechanism is replaced by a simplified feature aggregation strategy; and (3) `w/o 2D convolution`, where the back-end 2D convolutional reasoning module is removed. The full APTA2D model serves as the reference. Since the proposed framework is composed of front-end candidate-space compression and back-end relation reasoning, this ablation is designed to reveal how each module contributes to attribution accuracy and ranking quality. The model architecture and its three core components are described in Sectio III-B. (where pruning shrinks the candidate graph, multi-head attention highlights critical attack features, and 2D convolution captures higher-order spatial-semantic relations).

TABLE I
ABLATION STUDY RESULTS ON CCCS DATASET (MRR AND HITS@10).

Model	MRR	Hits@10
Full APTA2D	0.666	0.683
`w/o pruning`	0.341	0.664
`w/o multi-head attention`	0.336	0.657
`w/o 2D convolution`	0.188	0.504

Table I reports the ablation results. The full model achieves the best performance (MRR = 0.666, Hits@10 = 0.683). Removing pruning reduces performance to 0.341/0.664, demonstrating its importance for candidate filtering and top-rank prediction. Removing multi-head attention further degrades results to 0.336/0.657, indicating its role in discriminative feature weighting. The largest performance drop occurs without 2D convolution (MRR = 0.188, Hits@10 = 0.504), confirming its critical role in relational reasoning. Overall, the ablation study demonstrates that all three modules contribute to final performance, with APTA2D's superiority arising from their synergistic effect rather than any single component.

D. Validating the impact of attack knowledge

To validate the contribution of incorporating attack knowledge in enhancing attribution interpretability and reasoning performance, we compare the performance of APTA2D on the knowledge-focused dataset (CCCA) and the behavior-focused dataset (CCCS).

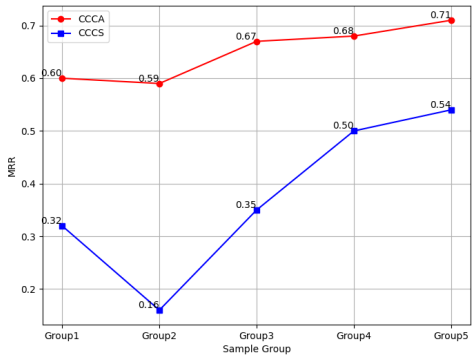

Fig. 4. Comparison under MRR

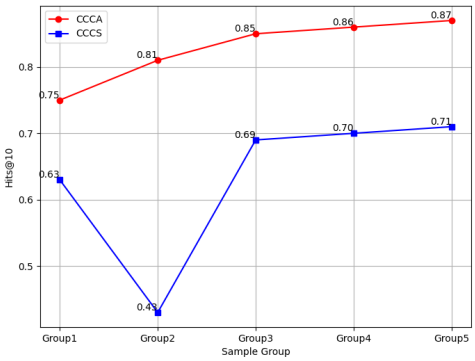

Fig. 5. Comparison under Hits@10

As shown in Figure 4 and Figure 5, the performance of APTA2D on the CCCA dataset (in terms of MRR and Hits@10) is significantly better than its performance on the CCCS dataset. For example, on the Group5 data, the MRR for CCCA is 0.71, while it is only 0.54 for CCCS. This indicates that integrating structured attack knowledge such as CVE, CWE, and CAPEC enables the model to better reveal the underlying logic and causal relationships among attack behaviors, thereby improving both the accuracy of reasoning and the interpretability of the results.

E. Application Results Analysis

To validate the attribution capability of APTA2D in real-world scenarios, we conducted an empirical analysis using the APT-C-01 (Poisonous Vine) attack as a case study. This attack involved multi-stage infiltration with multiple IP nodes. By inputting the collected IP list into the model, the algorithm successfully identified key nodes through iterative probability calculations. As shown in Figure 6, the model first identified an internal employee computer IP (132.54.xxx.xxx) as the initial entry point. Subsequently, it traced the correlation to an overseas C&C server (48.73.xxx.xxx) and finally located the source IP (53.62.xxx.xxx) in the attacker's region, achieving a high confidence score (>0.8). This process demonstrates the model's ability to reconstruct the complete attack chain from the point of entry to the ultimate source.

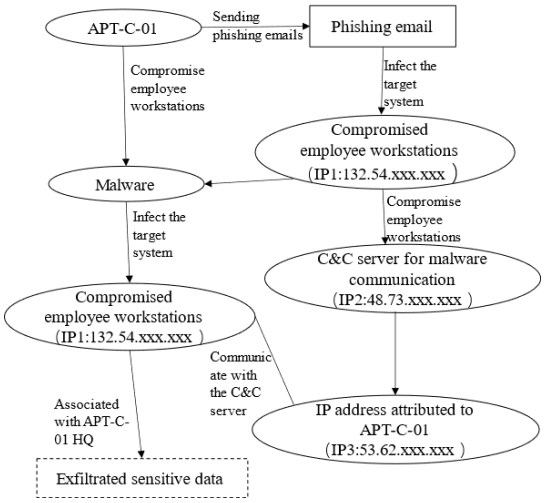

Fig. 6. A specific case study of APT-C-01 attack attribution.

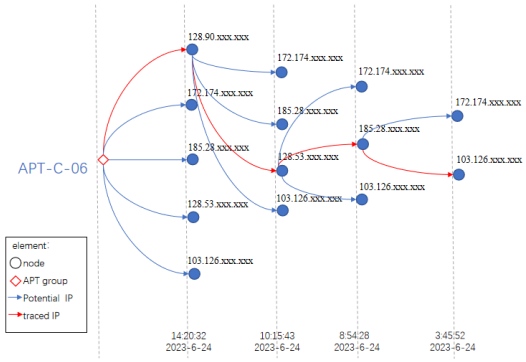

Fig. 7. A specific case study of APT-C-06 attack attribution.

Furthermore, to verify the generalizability of the proposed method across different attack scenarios, we tested it on APT-C-06 (Darkhotel) and APT-C-09 (Moha Grass). For APT-C-06, as illustrated in Figure 7, the model successfully traced the attack from the initial entry point (128.90.xxx.xx) through intermediate nodes to the source IP 103.126.xxx.xxx with a confidence of 0.93, recorded on June 24, 2023. For the APT-C-09 case, as depicted in Figure 8, the algorithm accurately backtracked through a complex five-stage path, ultimately identifying 74.119.xxx.xxx as the source IP on May 22, 2022. These results confirm that even when dealing with diverse and evolving attack chains, APTA2D maintains high precision and strong interpretability.

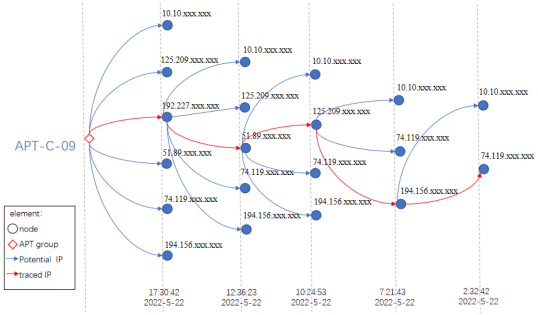

Fig. 8. A specific case study of APT-C-09 attack attribution.

V. Summary and Outlook

This paper proposes APTA2D, a coupled framework based on a "classify-first, reason-later" approach to address the interpretability and probabilistic reliance issues in existing APT tracing methods. The front-end utilizes a multi-head attention mechanism for probabilistic pruning to compress the candidate search space, while the back-end performs 2D convolutional reasoning on the condensed subgraph. This design reduces overall complexity from $O(N^2)$ to $O(\alpha^2 N^2)$ ($\alpha \leq 0.08$) and achieves multiplicative error decay. By integrating attack knowledge with behavioral features, APTA2D significantly enhances the accuracy of source IP localization and enables the complete reconstruction of attack chains.

Future work will address two identified limitations. First, the current model does not fully account for the dynamic temporal characteristics of multi-stage attacks. We plan to incorporate temporal modeling mechanisms and construct spatiotemporal fusion knowledge graphs to accurately capture the temporal evolution of attack chains. Second, the model's performance is limited in small-sample scenarios. We will explore data augmentation technologies to expand datasets and improve generalization capability in data-scarce environments.

References

[1] Stojanović B, Hofer-Schmitz K, Kleb U. APT datasets and attack modeling for automated detection methods: A review[J]. *Computers & Security*, 2020, 92: 101734.

[2] 360 Digital Security Group. Investigation Report on the Discovery of NSA Cyber Attacks Targeting Northwest Polytechnic University [EB/OL]. (2022-09-05) [2022-10-15]. https://mp.weixin.qq.com/s/0ReOzQMM5GS4xXRUPpKCvA.

[3] Kaspersky. Advanced Persistent Threats in 2021[EB/OL]. (2020-11-19) [2022-10-15]. https://www.kaspersky.com.cn/about/press-releases/2020_advanced-persistent-threats-in-2021-new-threat-angles-and-attack-strategy-changes-are-coming

[4] Berninger M. Going ATOMIC-Clustering and associating attacker activity at scale[J]. *FireEye Blog*, 2019.

[5] Li A. Research on Attack Organization Feature Extraction Method Based on Malware Genes[D]. Beijing University of Posts and Telecommunications, 2021. DOI: 10.26969/d.cnki.gbydu.2021.001346.

[6] Huang K, Lian Y, Feng D, Zhang H, Wu D, Ma X. A Graph Model-Based Method for Network Attack Attribution[J]. *Journal of Software*, 2022, 33(2): 683–698. DOI: 10.13328/j.cnki.jos.006314.

[7] Milajerdi S M, Eshete B, Gjomemo R, Venkatakrishnan V N. Poirot: Aligning attack behavior with kernel audit records for cyber threat hunting[C]. In: *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*. 2019: 1795–1812.

[8] Milajerdi S M, Gjomemo R, Eshete B, Sekar R, Venkatakrishnan V N. Holmes: Real-time APT detection through correlation of suspicious information flows[C]. In: *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2019: 1137–1152.

[9] Satvat K, Gjomemo R, Venkatakrishnan V N. Extractor: Extracting attack behavior from threat reports[C]. In: *2021 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, 2021: 598–615.

[10] Zhao J, Yan Q, Liu X, Li B, Zuo G. Cyber threat intelligence modeling based on heterogeneous graph convolutional network[C]. In: *23rd International Symposium on Research in Attacks, Intrusions and Defenses (RAID 2020)*. 2020: 241–256.

[11] Han X, Pasquier T, Bates A, Mickens J, Seltzer M. Unicorn: Runtime provenance-based detector for advanced persistent threats[J]. *arXiv preprint arXiv:2001.01525*, 2020.

[12] Wang Q, Hassan W U, Li D, Jee K, Yu X, Zou K, et al. You Are What You Do: Hunting Stealthy Malware via Data Provenance Analysis[C]. In: *NDSS*. 2020.

[13] Hassan W U, Bates A, Marino D. Tactical provenance analysis for endpoint detection and response systems[C]. In: *2020 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2020: 1172–1189.

[14] Alsaheel A, Nan Y, Ma S, Yu L, Walkup G, Celik Z B, et al. {ATLAS}: A sequence-based learning approach for attack investigation[C]. In: *30th USENIX Security Symposium (USENIX Security 21)*. 2021: 3005–3022.

[15] Cui M, Ma Y, Jiang Z, Ling Z, Li X. Hierarchical Attribution Analysis Technology for Network Attacks and Its Applications[J]. *Journal of Secrecy Science and Technology*, 2022(1): 24–30.

[16] Dong C, Lü M, Chen T, Zhu T. APT Attack Detection Method Based on Heterogeneous Provenance Graph Learning[J]. *Computer Science*, 2023, 50(4): 359–368.

[17] Han W, Xue J, Wang Y, Zhang F, Gao X. APTMalInsight: Identify and cognize APT malware based on system call information and ontology knowledge framework[J]. *Information Sciences*, 2021, 546: 633–664.

[18] Wang Q Q, Yan H B, Han Z H. Explainable APT attribution for malware using NLP techniques[C]. In: *Proceedings of the 21st IEEE International Conference on Software Quality, Reliability and Security (QRS)*. 2021: 70–80.

[19] Perry L, Shapira B, Puzis R. No-doubt: Attack attribution based on threat intelligence reports[C]. In: *2019 IEEE International Conference on Intelligence and Security Informatics (ISI)*. IEEE, 2019: 80–85.

[20] Abu M S, Selamat S R, Yusof R, Ariffin A. Formulation of Association Rule Mining (ARM) for an Effective Cyber Attack Attribution in Cyber Threat Intelligence (CTI)[J]. *International Journal of Advanced Computer Science and Applications*, 2021, 12(4).

[21] Hassan W U, Noureddine M A, Datta P, Bates A. OmegaLog: High-fidelity attack investigation via transparent multi-layer log analysis[C]. In: *Network and Distributed System Security Symposium (NDSS)*. 2020.

[22] Li T, Qiao W, Zhang J, Gao Y, Wang S, Shen Y, Ma J. Privacy-Preserving Attack Attribution Method Based on Graph Convolutional Neural Networks[J]. *Journal of Computer Research and Development*, 2021, 58(5): 1006–1020.

[23] Liang R, Gao Y, Zhao X. APT Attack Detection Method on Provenance Graphs Based on Sequence Feature Extraction[J]. *SCIENTIA SINICA Informationis*, 2022, 52(8): 1463–1480.

[24] Du Y, Ren W, Li W, Wang M, Wang W, Zhang H, Xia M. GA-ConvE: An APT attack prediction method based on combination of graph attention network and 2D convolution[J]. *Neural Networks*, 2025, 108216.

[25] Ren W, Zhang H, Lei Y. Network attack knowledge inference with graph convolutional networks and convolutional 2D KG embeddings[J]. *Scientific Reports*, 2025, 15(1): 34561.