ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

BÁO CÁO GIỮA KỲ

MÁY HỌC TRONG BẢO MẬT MẠNG VÀ HỆ THỐNG

TÊN ĐỀ TÀI: MAGIC: DETECTING ADVANCED PERSISTENT THREATS VIA MASKED GRAPH REPRESENTATION LEARNING

GIẢNG VIÊN HƯỚNG DẪN
TS. LÊ KIM HÙNG

HỌC VIÊN THỰC HIỆN

230202002 - TÔ THỊ MỸ ÂU
220202022 - NGUYỄN HỒNG SƠN
230202006 - NGÔ THÁI HƯNG

TP. HỒ CHÍ MINH, NĂM 2024

Mục lục

1 Giới thiệu tổng quan

Advance Persistent Threats là một quá trình tấn công mạng tinh vi trong thời gian dài, kẻ tấn công không bị phát hiện trong một thời gian dài và do đó, tiếp cận, đánh cắp, phá hoại dữ liệu quan trọng trong hệ thống. Nó sử dụng kết hợp nhiều công cụ và kỹ thuật phức tạp để xâm nhập và duy trì quyền truy cập tới hệ thống. Kẻ tấn công có mục tiêu dài hạn cụ thể (như gián điệp) và liên tục theo dõi và duy trì tương tác với hệ thống. Kẻ tấn công thường là tổ chức, có kỹ thuật và tổ chức tốt, được tài trợ thường xuyên, đôi khi do nhà nước tài trợ. Tóm lại, Advanced Persistent Threats (APTs) là những cuộc tấn công mạng kéo dài, tinh vi, có chủ đích rõ ràng và do những tổ chức lành nghề được tài trợ thực hiện [1]. Hầu hết các cuộc tấn công APT đều liên quan đến lỗ hổng zero-day và rất khó phát hiện.

Các nỗ lực nhằm phát hiện APT chủ yếu dựa vào các phương pháp: (1) xây dựng rules-base dựa trên các các mẫu APT phổ biến và so khớp với audit logs, (2) sử dụng thống kê các thành phần trong hệ thống: system entities, tương tác mạng... để phát hiện bất thường, (3) sử dụng các kỹ thuật học sâu để mô hình hóa tấn công APT hoặc hành vi hệ thống, sau đó phát hiện APT bằng cách phân loại hoặc phát hiện bất thường.

Các phương pháp trên đã chứng minh được sự hiệu quả đối với APT, nhưng còn một số thách thức: (1) học giám sát đòi hỏi phải có dữ liệu lớn về APT để train model, khi gặp APT mới thì khó phát hiện đây là tấn công, (2) các phương pháp thống kê không trích xuất được mối tương quan phức tạp trong log của hệ thống, (3) các phương pháp sử dụng DL có ưu điểm nhiều, nhưng yêu cầu quá nhiều tài nguyên tính toán, dẫn tới khó áp dụng trong thực tế.

2 Phân tích chi tiết bài báo

2.1 Giới thiệu

Các cuộc tấn công nâng cao và liên tục (APTs) ngày càng tinh vi được thực hiện bởi những kẻ tấn công có kỹ năng cao, gây ra mối đe dọa lớn cho cả các

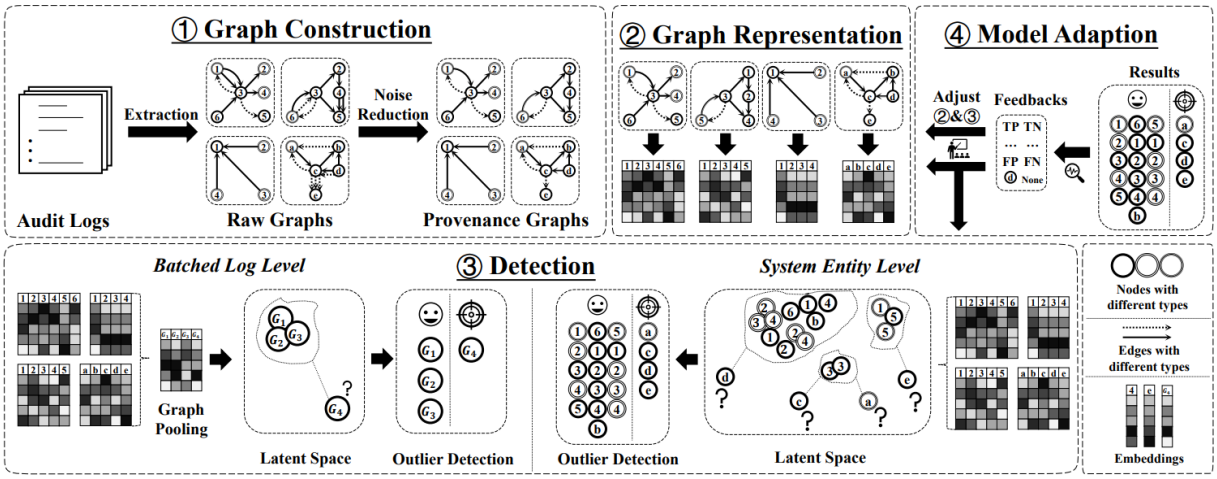
doanh nghiệp và tổ chức. Hiện nay, có nhiều phương pháp phát hiện APT dựa trên phân tích đồ thị nguồn gốc dữ liệu, nhằm trích xuất thông tin từ các bản ghi nhật ký và hỗ trợ phát hiện các mối đe dọa. Tuy nhiên, các phương pháp hiện tại vẫn còn gặp nhiều thách thức. Thiếu dữ liệu (LOD) trong các phương pháp có giám sát, yêu cầu kiến thức tiên nghiệm về APT, làm chúng dễ bị tấn công bởi các mối đe dọa mới. Tỷ lệ dương tính giả cao ở các phương pháp dựa trên thống kê, do không thể trích xuất ngữ nghĩa sâu từ các bản ghi. Chi phí tính toán lớn của các phương pháp học sâu, đặc biệt là các phương pháp dựa trên chuỗi và đồ thị, gây khó khăn trong việc triển khai thực tế.

Để giải quyết những vấn đề này, bài báo giới thiệu MAGIC, một phương pháp phát hiện APT tự giám sát mới. MAGIC kết hợp học biểu diễn đồ thị có che với các phương pháp phát hiện ngoại lệ đơn giản. MAGIC trước tiên xây dựng đồ thị từ các bản ghi nhật ký và sau đó sử dụng mô-đun auto-encoder đồ thị có che để tạo biểu diễn nhúng. Các biểu diễn này được phân tích bằng phương pháp phát hiện ngoại lệ không giám sát, giúp nhận diện các thực thể hệ thống bị tấn công. MAGIC được thiết kế với tính linh hoạt cao, cho phép phát hiện ở nhiều cấp độ khác nhau, từ phát hiện APT theo lô đến xác định đối tượng tấn công cụ thể. Nó có thể hoạt động hiệu quả trong cả môi trường không giám sát, bán giám sát, và giám sát hoàn toàn. Ngoài ra, MAGIC còn bao gồm một cơ chế thích nghi mô hình, giúp giảm tỷ lệ dương tính giả thông qua phản hồi từ người dùng MAGIC đã được đánh giá trên ba bộ dữ liệu APT khác nhau (DARPA Transparent Computing E3, StreamSpot, và Unicorn Wget), với các kết quả ấn tượng: độ chính xác 97,26%, độ nhạy 99,91%, và hiệu suất tính toán vượt trội so với các phương pháp hiện có, nhanh hơn ShadeWatcher tới 51 lần.

2.2 Phương Pháp

2.2.1 Tổng quan về MAGIC

MAGIC (Masked Graph Representation Learning for APT Detection) là một phương pháp học máy tự giám sát kết hợp với phát hiện ngoại lệ trên đồ thị nhằm mục đích phát hiện các mối đe dọa liên tục nâng cao (APTs). Các thành phần



Hình 2.1: Tổng qua về pipeline phát hiện bất thường của MAGIC

chính trong phương pháp của hệ thống này:

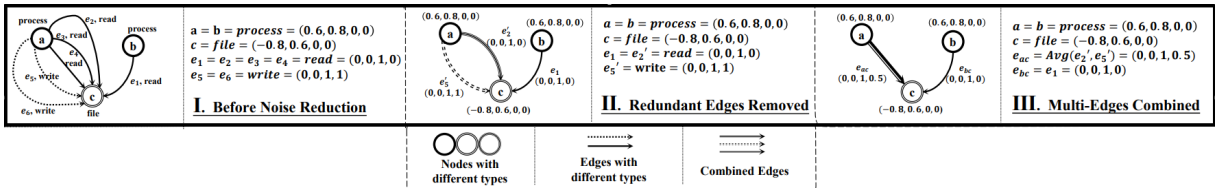
- **Xây dựng biểu đồ nguồn gốc (Provenance Graph Construction).**
- **Mô-đun học biểu diễn đồ thị (Graph Representation Module).**
- **Mô-đun phát hiện (Detection Module).**
- **Cơ chế thích nghi mô hình (Model Adaption Mechanism).**

2.2.2 Xây dựng biểu đồ nguồn gốc (Provenance Graph Construction)

Biểu đồ nguồn gốc được xây dựng từ các nhật ký kiểm toán hệ thống. Các nút và cạnh là các thực thể hệ thống và các tương tác trong các nhật ký:

- **Log Parsing:** Phân tích từng mục nhật ký để trích xuất các thực thể hệ thống (ví dụ: các tiến trình, tệp tin, luồng mạng,...) tương ứng là các nút và các tương tác giữa chúng (ví dụ: thực thi, đọc, kết nối,...) tương ứng là các cạnh của đồ thị.
- **Embedding ban đầu:** Gán các nhãn cho các nút và cạnh để biến chúng thành các vector đặc trưng.

- **Giảm nhiễu (Noise Reduction):** Để làm giảm độ phức tạp, MAGIC kết hợp các cạnh dư thừa giữa các cặp nút để tạo ra một cạnh duy nhất.



Hình 2.2: Ví dụ các bước tạo đồ thị của mô hình MAGIC

2.2.3 Mô-đun học biểu diễn đồ thị (Graph Representation Module)

Mô-đun học biểu diễn đồ thị của MAGIC nhằm tạo ra các vector biểu diễn hiệu quả cho các thực thể và trạng thái hệ thống:

- **Masked Graph Auto-Encoders:** Sử dụng bộ Auto-Encoders để học tự giám sát các biểu diễn :
  - **Feature Masking:** Các nút trong đồ thị sẽ được che ngẫu nhiên trong quá trình huấn luyện để mô hình học được các đặc trưng ẩn sâu trong biểu đồ mà không cần nhãn.
  - **Graph Encoder:** MAGIC sử dụng mạng attention trên đồ thị (GAT) để lan truyền và tổng hợp các đặc trưng từ các nút láng giềng.
  - **Graph Decoder:** Khôi phục lại các nút đã bị che để tối ưu hóa biểu diễn đồ thị.
- **Pooling and Embeddings:**
  - Sau khi đi qua các lớp GAT, MAGIC kết hợp các vector biểu diễn từ các nút để tạo ra một biểu diễn tổng thể của trạng thái hệ thống.

2.2.4 Mô-đun phát hiện ngoại lệ (Detection Module)

Dựa trên các biểu diễn đã học, MAGIC sử dụng mô-đun phát hiện ngoại lệ để phát hiện các thực thể hoặc trạng thái của hệ thống là bất thường:

- **Outlier Detection:** Khi có một thực thể mới, MAGIC tìm kiếm các láng giềng gần nhất của nó trong cây K-D và tính toán độ giống nhau (similarity) giữa chúng.
- **Anomaly Scoring:** MAGIC sử dụng khoảng cách Euclidean để đánh giá điểm ngoại lệ của một thực thể. Nếu điểm này vượt quá ngưỡng định trước, thực thể đó được đánh dấu là độc hại.
- **Batched Log Level Detection:** MAGIC sẽ cảnh báo nếu phát hiện thấy các trạng thái bất thường của lô nhật ký.
- **System Entity Level Detection:** Phát hiện các thực thể có hành vi bất thường.

2.2.5 Cơ chế thích nghi mô hình (Model Adaption Mechanism)

MAGIC có cơ chế thích nghi để đối phó với sự thay đổi của hệ thống (concept drift):

- **Feedback Learning:** Các kết quả phát hiện được xác nhận bởi người dùng sẽ được đưa lại vào MAGIC. Mô hình có thể học từ những hành vi hệ thống mới này và cải thiện khả năng phát hiện.
- **Memory Management:** Khi dữ liệu học từ feedbacks ngày càng nhiều, MAGIC sẽ loại bỏ các dữ liệu cũ để duy trì hiệu suất và giảm thiểu việc cảnh báo sai (false positives). Thông qua đó, MAGIC có thể thích nghi với các thay đổi hệ thống mới mà không bị giảm hiệu quả.

Table 1: Datasets for batched log level detection.

Dataset	Scenario	Malicious	#Log pieces	Avg. #Entity	Avg. #Interaction	Size(GB)
StreamSpot	CNN		100	8,989	294,903	0.9
	Download		100	8,830	310,814	1.0
	Gmail		100	6,826	37,382	0.1
	VGame		100	8,636	112,958	0.4
	YouTube		100	8,292	113,229	0.3
	Attack	✓	100	8,890	28,423	0.1
Unicorn Wget	Benign		125	265,424	975,226	64.0
	Attack	✓	25	257,156	949,887	12.6

Table 2: Datasets for system entity level detection.

Dataset	Scenario	Malicious	#Node	#Edge	Size (GB)
DARPA E3 Trace	Benign		3,220,594	4,080,457	15.40
	Extension Backdoor	✓	732		
	Pine Backdoor	✓	67,345		
	Phishing Executable	✓	5		
DARPA E3 THEIA	Benign		1,598,647	2,874,821	17.91
	Attack	✓	25,319		
DARPA E3 CADETS	Benign		1,614,189	3,303,264	18.38
	Attack	✓	12,846		

Hình 2.3: Các dataset thực nghiệm

Granularity	Dataset		Train Ratio	Ground Truth		#TP	#FP	#TN	#FN	Precision	Recall	FPR	F1-Score	AUC
				#Benign	#Malicious									
Batched log level	StreamSpot		80%	100	100	100.0	0.59	99.41	0.0	99.41%	100.00%	0.59%	99.71%	99.95%
	Unicorn Wget		80%	25	25	24.0	0.5	24.5	1.0	98.02%	96.00%	2.00%	96.98%	96.32%
System entity level	DARPA E3 Trace	All	80%	616,025	68,082	68,072	569	615,456	10	99.17%	99.98%	0.09%	99.57%	99.99%
		Extension Backdoor			732	727			5					
		Pine Backdoor			67,345	67,342			3					
		Phishing Executable			5	1			2					
	DARPA E3 THEIA	All	80%	319,448	25,319	25,318	456	318,992	1	98.23%	99.99%	0.14%	99.11%	99.87%
	DARPA E3 CADETS	All	80%	344,327	12,846	12,816	759	343,568	30	94.40%	99.77%	0.22%	97.01%	99.77%

Hình 2.4: Kết quả phát hiện của MAGIC trên các bộ dữ liệu khác nhau

Dataset	Approach	Train Ratio	Supervision	Precision	F1-Score	Recall	FPR
StreamSpot	StreamSpot	80%	B	73%	81%	91%	6.6%
	Unicorn (baseline)	75%	B	95%	96%	93%	1.6%
	Prov-Gem	80%	B,A	100%	97%	94%	**0%**
	ThreaTrace	75%	B	98%	99%	99%	0.4%
	MAGIC (Ours)	80%	B	99%	**99%**	**100%**	0.6%
Unicorn Wget	Unicorn (baseline)	80%	B	86%	90%	95%	15.5%
	Prov-Gem	80%	B,A	100%	89%	80%	**0%**
	ThreaTrace	80%	B	93%	95%	**98%**	7.4%
	MAGIC (Ours)	80%	B	98%	**97%**	96%	2.0%
DARPA E3 Trace	DeepLog	N/A	B,A	41%	51%	68%	2.7%
	Log2vec (baseline)	N/A	B,A	54%	64%	78%	1.8%
	ThreaTrace	N/A	B	72%	83%	99%	1.1%
	ShadeWatcher	80%	B,SA	97%	99%	**99%**	0.3%
	MAGIC (Ours)	80%	B	**99%**	**99%**	99%	**0.1%**
DARPA E3 THEIA	DeepLog	N/A	B,A	16%	15%	14%	0.5%
	Log2vec (baseline)	N/A	B,A	62%	64%	66%	0.3%
	ThreaTrace	N/A	B	87%	93%	99%	**0.1%**
	MAGIC (Ours)	80%	B	**98%**	**99%**	**99%**	0.1%
DARPA E3 CADETS	DeepLog	N/A	B,A	23%	35%	74%	4.4%
	Log2vec (baseline)	N/A	B,A	49%	62%	85%	1.6%
	ThreaTrace	N/A	B	90%	95%	**99%**	0.2%
	MAGIC (Ours)	80%	B	**94%**	**97%**	99%	**0.2%**

Hình 2.5: So sánh giữa MAGIC và các phương pháp phát hiện APT hiện đại trên các bộ dữ liệu khác nhau

2.3 Thực nghiệm và đánh giá của tác giả

Hình 2.3 giới thiệu ba bộ dữ liệu được sử dụng để đánh giá MAGIC. Dataset StreamSpot là một tập nhỏ gồm 600 lô nhật ký hệ thống, trong đó có một kịch bản tấn công. Dataset Unicorn Wget chứa 150 lô, trong đó có 25 lô với các cuộc tấn công chuỗi cung ứng khó phát hiện. Dataset DARPA E3, với hơn 51GB dữ liệu và hàng triệu thực thể hệ thống, đánh giá khả năng phát hiện của MAGIC trong môi trường phức tạp và quy mô lớn.

Kết quả cho thấy MAGIC phát hiện APT với độ chính xác cao trong nhiều kịch bản khác nhau như 2.4

MAGIC cho thấy hiệu quả cao trong việc phát hiện APT trên ba bộ dữ liệu khác nhau như hình 2.5. Cụ thể:

Phase	Component	Time consumption (s)		Peak Memory consumption (MB)
		with GPU	CPU only	
Graph Construction	N/A	642		2,610
Training	Graph Representation	151	685	1,564
	Detection	78		1,320
Inference	Graph Representation	5	10	2,108
	Detection	825		1,667

Hình 2.6: Chi phí hiệu suất của MAGIC trên tập dữ liệu con E3-Trace

Dataset	Train Ratio	Adaption	Test Ratio	FPR
StreamSpot	80%	N/A	20%	0.59%
Unicorn Wget	80%	N/A	20%	2.00%
DARPA E3 Trace	80%	N/A	20%	0.089%
	20%	N/A	20%	0.426%
	20%	20% FP	20%	0.272%
	20%	20% FP & TN	20%	0.220%
	20%	40% FP & TN	20%	0.173%

Hình 2.7: Tỷ lệ dương tính giả của MAGIC trên các tập dữ liệu khác nhau

Bộ dữ liệu StreamSpot: MAGIC đạt gần như kết quả hoàn hảo. MAGIC đạt độ chính xác trung bình 98.01% và độ hồi đáp 96.00%.

Bộ dữ liệu Unicorn Wget: Mặc dù khó phân biệt các cuộc tấn công ẩn nấp với hành vi lành tính, MAGIC vẫn phát hiện được 24/25 lô nhật ký tấn công với 0.5 dương tính giả.

Bộ dữ liệu DARPA E3: MAGIC đạt 99.91% độ hồi đáp và tỷ lệ dương tính giả 0.15%.

Kết quả cho thấy MAGIC vượt trội so với các phương pháp phát hiện APT hiện đại khác, như Unicorn, Prov-Gem, và ThreaTrace, về độ chính xác và hiệu suất.

MAGIC hoạt động với mức tiêu tốn tài nguyên tối thiểu, nhanh hơn nhiều lần so với các phương pháp tiên tiến nhất, giúp nó có thể áp dụng trong nhiều điều kiện khác nhau 2.6.

Bảng 2.7 trình bày kết quả thử nghiệm của MAGIC trên tập dữ liệu Trace, đặc biệt là hiệu quả của cơ chế điều chỉnh mô hình trong việc giảm số lượng false positives trong các bản ghi audit hợp lệ.

2.4 Kết luận của tác giả

MAGIC là một phương pháp phát hiện APT không cần giám sát, dựa trên mô hình hóa hành vi và phát hiện điểm bất thường. Áp dụng rộng rãi với chi phí tính toán thấp. Nó sử dụng học biểu diễn đồ thị để mô hình hóa hành vi hệ thống lành tính từ các nhật ký kiểm tra và phát hiện APT thông qua các phương pháp phát hiện ngoại lệ. Đánh giá trên ba bộ dữ liệu cho thấy MAGIC đạt kết quả tốt với tỷ lệ dương tính giả thấp và chi phí tính toán tối thiểu.

Tài liệu tham khảo

[1] Adel Alshamrani et al. “A survey on advanced persistent threats: Techniques, solutions, challenges, and research opportunities”. In: *IEEE Communications Surveys & Tutorials* 21.2 (2019), pp. 1851–1877.