

# ATOM: A Framework of Detecting Query-Based Model Extraction Attacks for Graph Neural Networks

Zhan Cheng  
University of Wisconsin, Madison  
Madison, Wisconsin, USA  
zcheng256@wisc.edu

Bolin Shen  
Florida State University  
Tallahassee, Florida, USA  
blshen@fsu.edu

Tianming Sha  
Arizona State University  
Tempe, Arizona, USA  
stianmin@asu.edu

Yuan Gao  
University of Wisconsin, Madison  
Madison, Wisconsin, USA  
ygao355@wisc.edu

Shibo Li  
Florida State University  
Tallahassee, Florida, USA  
sl24bp@fsu.edu

Yushun Dong  
Florida State University  
Tallahassee, Florida, USA  
yushun.dong@fsu.edu

## Abstract

Graph Neural Networks (GNNs) have gained traction in Graph-based Machine Learning as a Service (GMLaaS) platforms, yet they remain vulnerable to graph-based model extraction attacks (MEAs), where adversaries reconstruct surrogate models by querying the victim model. Existing defense mechanisms, such as watermarking and fingerprinting, suffer from poor real-time performance, susceptibility to evasion, or reliance on post-attack verification, making them inadequate for handling the dynamic characteristics of graph-based MEA variants. To address these limitations, we propose ATOM, a novel real-time MEA detection framework tailored for GNNs. ATOM integrates sequential modeling and reinforcement learning to dynamically detect evolving attack patterns, while leveraging  $k$ -core embedding to capture the structural properties, enhancing detection precision. Furthermore, we provide theoretical analysis to characterize query behaviors and optimize detection strategies. Extensive experiments on multiple real-world datasets demonstrate that ATOM outperforms existing approaches in detection performance, maintaining stable across different time steps, thereby offering a more effective defense mechanism for GMLaaS environments. Our source code is available at <https://github.com/LabRAI/ATOM>.

## Keywords

Graph Neural Networks, Model Extraction Attacks, Machine Learning as a Service, Security

## 1 Introduction

Graph Neural Networks (GNNs) [16, 33] have been widely studied for modeling graph-structured data, where nodes represent entities and edges capture their relationships. Accordingly, GNNs have also demonstrated promising performance in various real-world applications, such as financial fraud detection [23, 27], biomolecular

interaction analysis [1, 31], and personalized item recommendations [4, 8]. Despite its exceptional success, training GNNs has become increasingly costly due to the growing scale of both model and data. To democratize the access to powerful GNNs, Graph-based Machine Learning as a Service (GMLaaS) has emerged as a popular paradigm, which enables the model owner to provide easily accessible APIs for customers to use without disclosing the underlying GNN model. This facilitates the broader adoption of GNNs in various domains such as e-commerce [42], healthcare [7], and scientific research [6, 36, 51]. However, despite these advantages, GMLaaS platforms face significant security risks from model extraction attacks (MEAs) [20, 38]. These attacks allow adversaries to query a deployed API and systematically reconstruct a surrogate model that closely mimics the target model's behavior. Recent research [9] has demonstrated that MEAs pose a severe threat to GMLaaS platforms, which endanger both GMLaaS providers and users, leading to financial losses and potential downstream security threats. In the financial domain, for instance, service providers can deploy GMLaaS solutions to enhance credit card fraud detection [21, 47]. However, graph-based MEAs would enable adversaries to replicate fraud detection models, extract decision boundaries, and ultimately bypass fraud detection systems, thereby increasing the risk of large-scale financial crimes. Therefore, graph-based MEAs have emerged as a pressing security threat to GMLaaS platforms, highlighting the urgent need for robust defense mechanisms to mitigate these risks.

To counteract MEAs on GMLaaS, several mainstream defense strategies have been developed. A common defense strategy is watermarking, where model owners embed specially designed input-output patterns (as watermarks) into GNNs for ownership verification [3, 13, 18]. Specifically, given the specially designed input, if a certain GNN model produces the same patterns in its corresponding output, it is then implied that this GNN was obtained via MEA. While effective, watermarking may degrade model accuracy [19] and still leave the model vulnerable to attackers due to its passive nature. Another related approach is fingerprinting [40, 44], which aims to identify stolen models by comparing their outputs to a reference model [25, 26]. However, both fingerprinting and watermarking are passive rather than active and can only take effect after a GNN model has been stolen. More critically, none of these methods provides proactive detection—especially in GMLaaS, where queries are sequential, adaptive, and structurally dependent. This raises a

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).  
Conference'17, Washington, DC, USA

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.  
ACM ISBN 978-X-XXXX-XXXX-X/YY/MM  
<https://doi.org/XXXXXXX.XXXXXXX>

crucial question: How can we detect MEAs on GNNs proactively, rather than merely responding after an attack has already occurred? Although some DNN-based detection methods [22, 29, 37] could be adapted for GNNs, they often fail to capture the intricate relationships between nodes in graph-structured queries. As a result, attackers can bypass detection by leveraging node relationships and evolving their query strategies. In summary, while current defenses offer some level of post-attack verification, they lack the proactive capabilities needed to detect MEAs—especially in the context of GMLaaS. This gap highlights an urgent need for novel detection approaches that can monitor the adaptive and sequential queries tailored for specific graph structures.

Despite the critical importance need of proactively detecting MEAs on GMLaaS, it is a non-trivial task and we mainly face three fundamental challenges. (1) **Sequential Relationship of Graph-based MEA Queries.** In the strategical querying process, an attacker typically craft each query based on the historical information of the previous output sequence. Accordingly, the evolving trajectories of queries in the input space encodes key information identifying whether the user is malicious or legitimate. However, most existing approaches rely on the hypothesis that all queries are visible for the model provider to conduct defense [14], which thus makes it difficult to capture the query evolving patterns and flag potential attackers. (2) **Dynamic Characteristics of Graph-based MEA Variants.** In GMLaaS environments, adversaries could dynamically refine their attack strategies by exploiting the structural flexibility of graph-based queries. Rather than simply replicating well-known attack signatures [48, 52], they may adapt in real time, strategically avoiding high-risk nodes and targeting low-risk ones to evade detection. Thus, the second challenge is to design a detection framework that remains robust against evolving attack strategies. (3) **Necessity of considering multi-modal information.** Existing MEA detection methods, primarily designed for DNNs, often do not consider structural information. While effective in general cases, these methods may fail to capture the topological context of query-related nodes in GMLaaS. This is because graph-based queries involve both node attributes and topological information (e.g., multi-hop neighbors of a node). Thus, it is necessary to consider the information encoded in both modalities.

To address these challenges, we propose a novel framework, ATOM (Attacks deTector On GMLaaS), for real-time detection of MEAs targeting GMLaaS environments. Specifically, to tackle the first challenge, we introduce a differential query feature encoding mechanism that analyzes changes in query features across consecutive interactions. This approach enables our framework to adapt dynamically to evolving attack behaviors by continuously monitoring and evaluating incoming queries in real-time. Next, to address the second challenge, we refine our detection strategy through a reinforcement learning approach with a normalization factor, i.e., the Proximal Policy Optimization (PPO). This allows our detection policy dynamically adjust to evolving query patterns and reveal how attackers refine their methods. We further provide a theoretical analysis of these refinements. Finally, to overcome the third challenge, we enhance each query with values reflecting a node's structural importance, utilizing the  $k$ -core centrality. This enables the model to capture both local query traits and broader topological context, significantly improves its ability to distinguish between

legitimate and malicious queries, even in sparsely connected scenarios. Our main contributions can be summarized as follows:

- **Problem Formulation:** We provide a mathematical formulation of graph-based MEA detection in GMLaaS environments under the transductive setting, defining attack behaviors, detection objectives, and adversarial interactions.
- **Proposed Novel Framework:** To the best of our knowledge, ATOM is the first framework for proactive detection of graph-based MEAs in GMLaaS. Our empirical evaluations show that it outperforms existing methods adapted to this scenario.
- **Theoretical Analysis:** We conduct theoretical analysis on the query representation and derive formal bounds, offering a principled way to evaluate detection performance and optimize feature selection for adversarial query detection.

## 2 Preliminaries

In this section, we introduce the foundational concepts for detecting graph-based MEAs in a GMLaaS environment. The detection objective is to analyze user-submitted queries and determine whether the user is an attacker attempting to extract the deployed model. Our discussion covers the GMLaaS query-response framework, the GNN model, the objectives of both attackers and defenders and the formulation of the problem.

### 2.1 Graph-based Machine Learning as a Service

**Node-level Prediction Task.** GMLaaS systems provide a query-based interface that allows users to access pre-trained machine learning models hosted on cloud platforms [6]. Node-level prediction tasks typically operate under two primary learning paradigms: the transductive setting or the inductive setting [44]. In this work, we focus on the transductive setting, where the training graph used to train the GNN model is identical to the inference graph used for serving predictions and remains unchanged throughout the service. The GMLaaS system enables users to query node-level predictions while granting access to partial graph information.

**GMLaaS Query-Response Framework.** In the GMLaaS setting, each user  $u_i \in \mathcal{U}$  submits a query  $q_{i,t}$  targeting a specific node  $v_{i,t} \in \mathcal{V}$  within a graph  $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{X})$ , where  $\mathcal{V}$  represents the set of nodes,  $\mathcal{E}$  denotes the set of edges, and  $\mathbf{X} \in \mathbb{R}^{|\mathcal{V}| \times d}$  represents the node feature matrix. Upon receiving the query, the GMLaaS system provides a predicted label, denoted as  $y_{i,t} = \mathcal{M}(v_{i,t})$ , where  $y_{i,t} \in \mathcal{C}$ , and  $\mathcal{C}$  is the set of possible class labels.

Additionally, user  $u_i$  can access the one-hop subgraph  $\mathcal{G}_{i,t} = (\mathcal{V}_{i,t}, \mathcal{E}_{i,t}, \mathbf{X}_{i,t})$  centered around the queried node  $q_{i,t}$ . Here,  $\mathcal{V}_{i,t} = \{v_{i,t}\} \cup \{w \mid (w, v_{i,t}) \in \mathcal{E}\}$  includes  $q_{i,t}$  and its one-hop neighbors,  $\mathcal{E}_{i,t} = \{(w, v) \in \mathcal{E} \mid w, v \in \mathcal{V}_{i,t}\}$  contains the edges connecting nodes in  $\mathcal{V}_{i,t}$ , and  $\mathbf{X}_{i,t}$  represents features of nodes in  $\mathcal{V}_{i,t}$ .

**User and Query Sequences.** Consider a set of users denoted as  $\mathcal{U} = \{u_1, u_2, \dots, u_M\}$ , where each user submits queries independently. The query history of a user  $u_i \in \mathcal{U}$  is represented as a query sequence  $\mathbf{Q}_i = \{q_{i,1}, q_{i,2}, \dots, q_{i,T_i}\}$ , where  $T_i$  denotes the total number of queries made by  $u_i$ . Since queries arrive sequentially,  $T_i$  also represents the total time steps of queries for user  $u_i$ .

## 2.2 Graph Neural Networks (GNNs)

Acting as the backbone of our proposed framework, a Graph Neural Network (GNN)  $\mathcal{M}$  is trained on the static graph  $\mathcal{G}$  for a specific downstream learning task. The basic operation of GNN between  $l$ -th and  $(l+1)$ -th layer can be formulated as follows:

$$\mathbf{h}_v^{(l+1)} = \sigma(\text{COMBINE}(\mathbf{h}_v^{(l)}, f(\{\mathbf{h}_u^{(l)} : u \in \mathcal{N}(v)\}))), \quad (1)$$

where  $\mathbf{h}_v^{(l+1)}$  and  $\mathbf{h}_v^{(l)}$  represent the embedding of node  $v$  at  $l$ -th and  $(l+1)$ -th layer correspondingly. The node feature matrix  $\mathbf{X}$  serves as the input to the GNN, where each node feature  $\mathbf{x}_v$  initializes the corresponding hidden representation  $\mathbf{h}_v^{(0)}$ . Given the adjacency matrix  $\mathbf{A}$ , the neighbor set of node  $v$  is denoted as  $\mathcal{N}(v)$ . The aggregation function  $f(\cdot)$  gathers information from the neighbors of  $v$ , and the combining function  $\text{COMBINE}(\cdot)$  integrates this information with the current hidden representation  $\mathbf{h}_v^{(l)}$ . An activation function  $\sigma(\cdot)$  (e.g., ReLU), is applied to introduce non-linearity. Given the output of the last GNN layer by matrix  $\mathbf{Z} \in \mathbb{R}^{n \times c}$ , the prediction  $\hat{\mathbf{Y}}$  of GNN can be written as  $\text{softmax}(\mathbf{Z}) \in \mathbb{R}^{n \times c}$  for node classification, and  $\text{sigmoid}(\mathbf{Z}^T \mathbf{Z}) \in \mathbb{R}^{n \times n}$  for link prediction [17].

## 2.3 Adversary's Objective

The adversary's goal is to reconstruct a surrogate model  $\mathcal{M}'$  that closely approximates the behavior of the victim GNN model  $\mathcal{M}$ . This is achieved by systematically querying the GMLaaS system and collecting query-response pairs.

**Adversary's Knowledge.** Following the attack taxonomy in [43], we assume that the attacker possesses partial knowledge of the graph's structure and attributes. For instance, in a social network system, an attacker may access partial user connections and attributes through public profiles. Specifically, the attacker  $u_i$  can access a subgraph  $\mathcal{G}' \subset \mathcal{G}$ . At time step  $t$ ,  $u_i$  submits query  $q_{i,t}$  to access the one-hop subgraph  $\mathcal{G}_{i,t} \subset \mathcal{G}'$  and the node features  $\mathbf{X}_{i,t}$ , where  $\mathcal{G}_{i,t} = (\mathcal{V}_{i,t}, \mathcal{E}_{i,t}, \mathbf{X}_{i,t})$ . The attacker can dynamically refine their query strategy based on information obtained from prior queries.

**Extracted Model Training.** The attacker trains the extracted model  $\mathcal{M}'$  by minimizing the prediction error between the victim model  $\mathcal{M}$  and  $\mathcal{M}'$ . This objective is formulated as:

$$\min_{\mathcal{M}'} \mathbb{E}_{v \in V} [\mathcal{L}(\mathcal{M}(v), \mathcal{M}'(v))], \quad (2)$$

where  $\mathcal{L}$  is the loss function measuring the prediction difference.

## 2.4 Defender's Objective

The defender's goal is to detect adversarial users by classifying users based on their query sequences. This requires designing a detection function  $Z$ , which assigns a classification label:  $d_{i,T} = Z(\mathbf{Q}_{i,T})$ , where  $d_{i,T} \in \{0, 1\}$ , with 0 representing a legitimate user and 1 representing an attacker. Formally, the defender aims to learn an optimal function  $Z^*$  that maximizes detection accuracy while minimizing false positives and false negatives:

$$Z^* = \arg \max_Z \mathbb{E}[\mathbb{I}(Z(\mathbf{Q}_{i,T}) = y_i)], \quad (3)$$

where  $y_i$  is the truth label of user  $u_i$ ,  $\mathbb{I}(\cdot)$  is the indicator function.

## 2.5 Problem Statement

**PROBLEM 1. Graph-based MEA detection in GMLaaS environments under the transductive setting.** Let  $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{X})$  be a static attributed graph. A GNN  $\mathcal{M}$  is deployed on an GMLaaS platform under the transductive setting. Each user  $u_i \in \mathcal{U}$  submits a query sequence  $\mathbf{Q}_i = \{q_{i,t}\}_{t=1}^{T_i}$ . Our goal is to design a detection function  $Z$  that assigns a label  $d_{i,T}$  to user  $u_i$  based on their query sequence  $\mathbf{Q}_{i,T}$  up to time step  $T$ , aiming to maximize the expected classification accuracy:  $Z^* = \arg \max_Z \mathbb{E}[\mathbb{I}(Z(\mathbf{Q}_{i,T}) = y_i)]$ , so that attackers and legitimate users can be accurately classified.

## 3 Methodology

### 3.1 Framework Overview

An overview of the proposed framework is shown in Figure 1. Specifically, it consists of two modules: (1) **Attack Simulation**. This module generates realistic model extraction attack sequences to serve as training data for the detection model. To achieve this, we integrate active learning techniques to mimic adversarial query behaviors under realistic GMLaaS constraints. (2) **Attack Detection**. This module consists of query embedding, a sequential network, and a reinforcement learning-based detection mechanism. It processes query sequences and classifies users as attackers or legitimate users based on their query behaviors.

### 3.2 Attack Simulation

A major challenge in constructing a reliable detection mechanism is obtaining high-fidelity training data that accurately represent real-world attacker behaviors. Instead of relying on passive observation, we proactively simulate realistic MEAs through the following steps.

**3.2.1 Active learning based Attacks.** Since existing Graph-based MEAs are relatively limited, we adapt active learning (AL) [35] to construct realistic attack query sequences, since both AL and MEAs share a common objective: maximizing knowledge extraction from a model while operating under strict query constraints. We simulate attacks using three representative algorithms:

**AGE [2]** (Active Exploration-Based Query Strategy). At each time step  $T$ , AGE selects a node  $v_T$  based on a scoring function  $S(v_T)$ , which integrates: Information entropy (uncertainty), Information density (node importance), and Graph centrality (network influence). To align with GMLaaS constraints, we modify AGE with the average highest score within one-hop subgraph of  $v_T$ , defined as:

$$S_{avg}(v_T) = \frac{\sum_{v \in V_T} S(v) + S(v_T)}{V_T + 1}, \quad (4)$$

This ensures that query sequences reflect real-world constraints on node accessibility. The generated query sequence follows a descending order based on  $S_{avg}$ .

**GRAIN [50]** (Influence Maximization-Based Query Strategy). At each time step  $T$ , GRAIN selects a node  $v_T$  to maximize the score function  $S(\mathcal{G}'_s)$ , where:

$$S(\mathcal{G}'_s) = \frac{|\sigma(\mathcal{G}'_s)|}{|\hat{\sigma}|} + \gamma \frac{D(\mathcal{G}'_s)}{\hat{D}}. \quad (5)$$

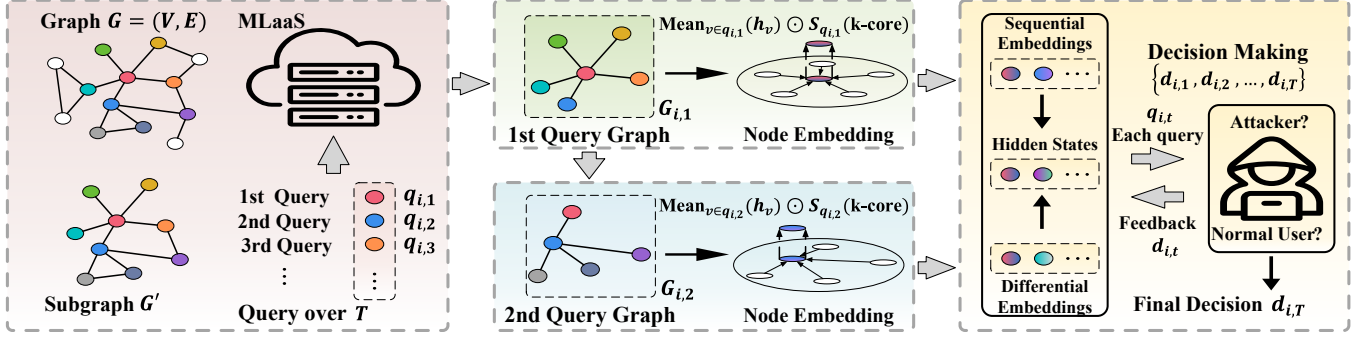


Figure 1: An illustration of the framework with the query behavior and the detection mechanism.

Here,  $\sigma(\mathcal{G}'_s)$  represents the influence spread of the selected subgraph  $\mathcal{G}'_s$ , and  $D(\mathcal{G}'_s)$  measures query diversity. The generated query sequence follows a descending order based on  $S(\mathcal{G}'_s)$ .

**IGP [49]** (Label-Informed Query Strategy). At each time step  $T$ , IGP selects the next node  $v_T$  by assuming the pseudo-label with the highest confidence in its softmax output  $\hat{y}_T$ , thereby maximizing the entropy change in its neighborhood. To improve efficiency, we first pre-filter nodes using a ranking score:

$$s = \alpha \cdot \mathcal{P}_{centrality} + (1 - \alpha) \cdot \mathcal{P}_{entropy}, \quad (6)$$

Only the top-ranked nodes are selected for querying, reducing query overhead while maximizing model information extraction.

**3.2.2 Query Sequence Generation.** We utilize the above attack simulation strategies to train surrogate models  $\mathcal{M}'$  with corresponding query sequences  $\mathcal{Q}_{\mathcal{M}'}$ . However, not all extracted sequences are considered valid attacks. We apply a quality threshold  $F_{\text{threshold}}$ , retaining only high-fidelity attack sequences:

$$\mathcal{Q}_{\text{attack}} = \{\mathcal{Q}_{\mathcal{M}'_1}, \mathcal{Q}_{\mathcal{M}'_2}, \dots, \mathcal{Q}_{\mathcal{M}'_H}\}, \quad (7)$$

where  $F(\mathcal{M}'_j) > F_{\text{threshold}}$ . For training balance, we also include legitimate user query sequences:

$$\mathcal{Q}_{\text{normal}} = \{\mathcal{Q}_1, \mathcal{Q}_2, \dots, \mathcal{Q}_N\} \quad (8)$$

All sequences are labeled (attack = 1, normal = 0), shuffled, and assigned to a set of users  $\mathcal{U}$ , where  $|\mathcal{U}| = |\mathcal{Q}_{\text{attack}}| + |\mathcal{Q}_{\text{normal}}|$ . Thus, for each user  $u_i \in \mathcal{U}$ , a query sequence  $\mathcal{Q}_i = \{q_{i,1}, q_{i,2}, \dots, q_{i,T_i}\}$  is generated for training the attack detection model.

### 3.3 Attack Detection

Attack detection in GMLaaS environments is more than a binary classification problem. Real attackers adapt over time steps, modifying their queries based on model responses to evade detection. A detection system that classifies queries individually, without considering their sequential nature or strategic dependencies, is insufficient. Furthermore, the detection mechanism must be resilient, continuously refining its strategy as attack patterns evolve.

**3.3.1 Sequences Embedding.** At time step  $T$ , each query  $q_{i,T}$  is transformed into an embedding  $h_{i,T}$ , incorporating both node features and graph topological information:

$$h_{i,T} = \frac{\sum_{v \in \mathcal{V}_{i,T}} h_v}{|\mathcal{V}_{i,T}|} \odot S\left(\frac{\log(p_{v_{i,T}})}{\log(p_{\max})}\right), \quad (9)$$

where  $h_v$  represents node embeddings obtained from the GMLaaS model  $\mathcal{M}$ ,  $p_{v_{i,T}}$  is the  $k$ -core value of the central node  $v_{i,T}$ , and  $p_{\max}$  is the maximum  $k$ -core value in graph  $\mathcal{G}$ . Here,  $S(x)$  is a scaling function defined as:

$$S(x) = 1 + \lambda \cdot (\sigma(\lambda \cdot x) - 0.5) \times 2, \quad (10)$$

where  $\lambda$  is a hyperparameter controlling the effect of topological scaling, ensuring that  $h_{i,T}$  is modulated based on graph structure while keeping variations within  $[1 - \lambda, 1 + \lambda]$ . This embedding mechanism ensures that detection captures structural dependencies, making it harder for adversaries to exploit low-connectivity nodes for stealthy model extraction.

**3.3.2 Sequential Modeling.** Model extraction attacks evolve over time steps—each query is part of a larger, strategic attack sequence. To capture temporal dependencies, we enhance a classic Gated Recurrent Unit (GRU) [5] with: (1) Differential input encoding, which highlights query-to-query variations (2) A fusion gate, selectively incorporating past and present query features. (3) A mapping matrix, adjusting hidden states based on past classification decisions. At time step  $T$ , we compute the differential input  $\delta_{i,T}$  as  $\delta_{i,T} = h_{i,T} - h_{i,T-1}$ , where  $h_{i,0} = 0$ . We introduce a fusion gate  $g_T$ , which determines how much of the current query embedding  $h_{i,T}$  and its differential input  $\delta_{i,T}$  should be retained:

$$g_T = \sigma(\mathbf{W}_g \cdot \text{Concat}(\delta_{i,T}, h_{i,T}) + b_g), \quad (11)$$

where  $\mathbf{W}_g$  and  $b_g$  are learnable parameters. The input is given by:

$$x_T = g_T \odot \delta_{i,T} + (1 - g_T) \odot h_{i,T}. \quad (12)$$

This ensures that detection is based on the "story" behind a sequence of queries, rather than treating them as isolated requests.

The sequential hidden state is updated with the GRU mechanism:

$$h_{i,T}^{\text{seq}} = [(1 - z_T) \odot h_{i,T-1}^{\text{seq}} + z_T \odot \tilde{h}_T]^T \cdot \mathbf{m}_{i,T-1}, \quad (13)$$

where  $z_T$  is the update gate,  $\tilde{h}_T$  is the candidate state, and  $\mathbf{m}_{i,T-1}$  is a mapping matrix introduced to adjust the hidden state based on

historical classification actions. Here, the mapping matrix  $\mathbf{m}_{i,T-1}$  is computed as:

$$\mathbf{m}_{i,T-1} = \mathbf{W}_a \cdot \mathbf{p}_{d_{i,T-1}} + \mathbf{b}_a, \quad (14)$$

where  $\mathbf{p}_{d_{i,T-1}}$  represents the classification probabilities from the PPO-based reinforcement learning module, and  $\mathbf{W}_a, \mathbf{b}_a$  are learnable transformation matrices. This adjustment ensures that past classification decisions influence future query analysis, making detection more adaptive to evolving attack strategies.

**3.3.3 Decision Making.** Static detection rules cannot adapt to emerging attack strategies. To enable continuous learning, we integrate reinforcement learning (RL) via Proximal Policy Optimization (PPO) [34]. At time step  $T$ , the system observes a state  $s_{i,T}$ , selects an action  $d_{i,T} \in \{0, 1\}$  (attacker or legitimate user), and receives a reward  $R(s_{i,T}, d_{i,T})$  based on classification correctness:

$$R(s_{i,T}, d_{i,T}) = \begin{cases} R_w(s_{i,T}, d_{i,T}), \\ R_{\text{penalty}} & \text{for classification bias,} \end{cases} \quad (15)$$

where  $R_w(s_{i,T}, d_{i,T})$  is defined as:

$$R_w(s_{i,T}, d_{i,T}) = \begin{cases} w_{\text{TP}}, & \text{if } d_{i,T} = 1 \text{ and } l = 1, \\ w_{\text{TN}}, & \text{if } d_{i,T} = 0 \text{ and } l = 0, \\ -w_{\text{FN}}, & \text{if } d_{i,T} = 0 \text{ and } l = 1, \\ -w_{\text{FP}}, & \text{if } d_{i,T} = 1 \text{ and } l = 0, \end{cases} \quad (16)$$

The bias penalty  $R_{\text{penalty}}$  is applied when the model overwhelmingly classifies users as attackers or normal users, defined as:

$$R_{\text{penalty}} = -p, \quad \text{where } p > w_{\text{FN}} > \max\{w_{\text{TP}}, w_{\text{TN}}, w_{\text{FP}}\} \quad (17)$$

## 4 Theoretical Analysis

In this section, we establish the theoretical foundation of our proposed framework by linking the graph-based query interaction scenario to fundamental mathematical concepts. Specifically, we model user behavior as a dominating set problem on a weighted graph, where the objective is to balance coverage and weight minimization. In this setting, legitimate users seek to maximize coverage efficiently, whereas attackers attempt to maximize the total weight of accessed nodes while minimizing coverage to evade detection. To address this challenge, we demonstrate that incorporating first-order and second-order differences in query embeddings is crucial for capturing adversarial behaviors, particularly in dynamic query sequences. Additionally, we provide a probabilistic interpretation of ATOM in the appendix.

### 4.1 Query as a Dominating Set Problem

In this subsection, we interpret the process of accessing a subgraph by a user as constructing a dominating set. The objective for a normal user is to cover necessary nodes while minimizing resource costs, typically quantified by node weights. However, adversarial users often follow a different strategy: they attempt to maximize the total weight of accessed nodes while keeping the coverage rate low to remain undetected. Theorem 4.1 formalizes this trade-off.

**THEOREM 4.1.** *Consider a covering graph  $\mathcal{D}$  in the graph  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ , aiming to cover at least  $\beta \in [0, 1]$  percent nodes of  $\mathcal{G}$ , while minimizing  $\sum_{u \in \mathcal{A}} w(u)$ , where  $w(u)$  is the weight of node  $u$  and*

*$\mathcal{A}$  represents the set of nodes not being covered, then the maximum covering percentage is given by*

$$\beta \leq \min\left\{\frac{|\mathcal{D}| - \frac{W}{w_{\mathcal{A}}}}{\frac{n}{\delta} - \frac{W}{w_{\mathcal{A}}}}, \frac{|\mathcal{D}| \cdot \delta}{n}\right\}. \quad (18)$$

*Here,  $|\mathcal{D}|$  represents the number of nodes in  $\mathcal{D}$ ,  $W$  is the total weight in  $\mathcal{G}$ ,  $w_{\mathcal{A}}$  represents the average weight in  $\mathcal{A}$  and  $\delta$  is the smallest degree for nodes in  $\mathcal{D}$ .*

Our results indicate that increasing the minimum degree of the covered subgraph while querying would enhance the coverage, which could be adopted to implement a high-quality MEA. Based on this observation, we integrate  $k$ -core values into the query embeddings to prioritize structurally significant nodes. This ensures that the detected subgraphs remain well-connected, thereby constraining the attacker's ability to manipulate coverage.

### 4.2 Incremental Changes in Query Behavior

In this subsection, we aim to model how queries change over time steps. Specifically, we examine incremental changes in graph coverage and node weights. We define the first-order difference to measure how the weight of uncovered nodes evolves as new nodes are queried, capturing gradual shifts in user behavior. Proposition 4.2 relates the weight reduction per node to the change in coverage.

**PROPOSITION 4.2.** *Consider a changing graph  $\mathcal{D}_{t-1}$  and  $\mathcal{D}_t$  in the graph  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ , where  $\mathcal{D}_{t-1} \subset \mathcal{D}_t \subset \mathcal{G}$ , achieving at least  $\beta_{t-1}$  covering rate with the lowest degree  $\delta_{t-1}$  and at most  $\beta_t$  covering rate with the lowest degree  $\delta_t$ , respectively. Also, suppose that  $\mathcal{A}_{t-1}$  and  $\mathcal{A}_t$  represent the set of nodes that are not covered, with the corresponding weight  $W_{\mathcal{A}_{t-1}}$ ,  $W_{\mathcal{A}_t}$  and the average weight  $w_{\mathcal{A}_t}$ ,  $w_{\mathcal{A}_{t-1}}$ . We then get*

$$\frac{\Delta W_{\mathcal{A}}}{\Delta |\mathcal{D}|} \leq \frac{(\beta_t - \beta_{t-1}) \cdot W}{|\mathcal{D}_{t-1}|(1 - \frac{\delta_{t-1}}{\delta_t})}, \quad (19)$$

*and  $\delta_t > \delta_{t-1}$ . Here,  $|\mathcal{D}_{t-1}|, |\mathcal{D}_t|$  represents the number of nodes in  $\mathcal{D}_{t-1}, \mathcal{D}_t$ ,  $W$  is the total weight in  $\mathcal{G}$ .*

Here,  $\frac{\Delta W_{\mathcal{A}}}{\Delta |\mathcal{D}|}$  acts as a first-order difference, quantifying how the weight of uncovered nodes evolves as new nodes are queried. This provides a direct measure of the trade-off between weight minimization and coverage expansion, enabling the model to capture gradual shifts in adversarial behavior.

However, first-order differences alone may fail when attackers target high-weight nodes while minimizing coverage expansion. In such cases, the weight reduction per added node may fluctuate, making it necessary to consider second-order differences to capture variations in how these changes occur over time steps.

### 4.3 Strategy Shifts in Query Behavior

To capture fluctuations in the rate of coverage expansion and weight reduction discussed above, we introduce the second-order differences to measure the change in first-order differences. Specifically, we aim to address the challenge when attackers adjust their query strategy by alternating between targeting high-weight nodes and optimizing coverage. The second-order difference can reveal these shifts, while the first-order difference may appear stable in this scenario. Proposition 4.3 demonstrates that it serves as a key indicator

of irregular access patterns. The second-order difference in weight is given as follows.

**PROPOSITION 4.3.** *Consider changing graphs  $\mathcal{D}_{t-1}$ ,  $\mathcal{D}_t$  and  $\mathcal{D}_{t+1}$  in the graph  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ , where  $\mathcal{D}_{t-1} \subset \mathcal{D}_t \subset \mathcal{D}_{t+1} \subset \mathcal{G}$ , achieving at least  $\beta_{t-1}$  covering rate with the lowest degree  $\delta_{t-1}$  and at most  $\beta_t$  covering rate with the lowest degree  $\delta_t$ . Also, there is at least  $\beta'_t$  covering rate and at most  $\beta_{t+1}$  covering rate at time step  $t$  and  $t+1$ . Suppose that  $\mathcal{A}_{t-1}$ ,  $\mathcal{A}_t$  and  $\mathcal{A}_{t+1}$  represent the set of nodes that are not covered, with the corresponding weight  $W_{\mathcal{A}_{t-1}}$ ,  $W_{\mathcal{A}_t}$ ,  $W_{\mathcal{A}_{t+1}}$  and the average weight  $w_{\mathcal{A}_{t-1}}$ ,  $w_{\mathcal{A}_t}$ ,  $w_{\mathcal{A}_{t+1}}$ . We then get*

$$\frac{\Delta^2 W_{\mathcal{A}}}{\Delta^2 |\mathcal{D}|} \geq \left| \frac{W_d}{n} \frac{\Delta \delta_t}{\Delta \beta_t} - \frac{W \delta_{t+1}}{|\mathcal{D}_t|} \frac{\Delta \beta_{t+1}}{\Delta \delta_{t+1}} \right|, \quad (20)$$

and  $\delta_{t+1} > \delta_t > \delta_{t-1}$ . Here,  $|\mathcal{D}_{t-1}|$ ,  $|\mathcal{D}_t|$  and  $|\mathcal{D}_{t+1}|$  represent the number of nodes in  $\mathcal{D}_{t-1}$ ,  $\mathcal{D}_t$  and  $\mathcal{D}_{t+1}$ ,  $W$  is the total weight in  $\mathcal{G}$ , assuming  $W_{\mathcal{A}_{t-1}} - W_{\mathcal{A}_t} \geq W_d$ .

To integrate this into our framework, we define the second-order difference in query embeddings as  $\delta_{i,T} = h_{i,T} - h_{i,T-1}$ , which captures temporal variations. These help identify adversarial patterns where attackers subtly shift their behavior to avoid detection.

#### 4.4 Importance of Second-Order Differences

To present the importance of introducing second-order differences, we establish a condition in Theorem 4.4 when the second-order differences contribute more significantly to detection than the first-order ones. Specifically, our analysis derives a threshold. If the traditional detection mechanism could exceed this threshold, we say the second-order differences are well worth being considered.

**THEOREM 4.4.** *Consider changing graphs  $\mathcal{D}_{t-1}$ ,  $\mathcal{D}_t$  and  $\mathcal{D}_{t+1}$  with graph  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ , where  $\mathcal{D}_{t-1} \subset \mathcal{D}_t \subset \mathcal{D}_{t+1} \subset \mathcal{G}$ , aiming at achieving at least  $\beta_{t-1}$  covering rate with the lowest degree  $\delta_{t-1}$  and at most  $\beta_t$  covering rate with the lowest degree  $\delta_t$ . Also, there is at least  $\beta'_t$  covering rate and at most  $\beta_{t+1}$  covering rate at time step  $t$  and  $t+1$ . Suppose that  $\mathcal{A}_{t-1}$ ,  $\mathcal{A}_t$  and  $\mathcal{A}_{t+1}$  represent the set of nodes that are not covered, with the corresponding weight  $W_{\mathcal{A}_{t-1}}$ ,  $W_{\mathcal{A}_t}$ ,  $W_{\mathcal{A}_{t+1}}$  and the average weight  $w_{\mathcal{A}_{t-1}}$ ,  $w_{\mathcal{A}_t}$ ,  $w_{\mathcal{A}_{t+1}}$ . The second-order differences become essential, that is,  $\frac{\Delta^2 W_{\mathcal{A}}}{\Delta^2 |\mathcal{D}|} \geq \frac{\Delta W_{\mathcal{A}}}{\Delta |\mathcal{D}|}$  holds when*

$$W_d \geq \frac{n \delta_{t+1} \Delta \beta_t}{|\mathcal{D}_{t-1}| \Delta \delta_t} \left( \frac{\Delta \beta_t}{\Delta \delta_t} + \frac{\Delta \beta_{t+1}}{\Delta \delta_{t+1}} \right) W, \quad (21)$$

and  $\delta_{t+1} > \delta_t > \delta_{t-1}$ . Here,  $|\mathcal{D}_{t-1}|$ ,  $|\mathcal{D}_t|$  and  $|\mathcal{D}_{t+1}|$  represent the number of nodes in  $\mathcal{D}_{t-1}$ ,  $\mathcal{D}_t$  and  $\mathcal{D}_{t+1}$ ,  $W$  is the total weight in  $\mathcal{G}$ , assuming  $W_{\mathcal{A}_{t-1}} - W_{\mathcal{A}_t} \geq W_d$ .

This insight highlights the importance of dynamically adjusting detection strategies based on the observed query behavior. To leverage this, we incorporate a fusion gate  $g_T$  that adaptively balances the first and second-order differences. This ensures that the model remains robust against evolving attack strategies, adjusting its detection focus as needed. We prove this theorem by empirical evaluations in section 5.3 and section 5.4.

## 5 Experimental Evaluations

We conduct a series of experiments to evaluate the performance of the proposed framework. Specifically, we seek to address the

following research questions: **RQ1:** How effectively can the proposed model capture attacks compared to baseline methods? **RQ2:** How do the individual components contribute to the overall performance of the proposed model? **RQ3:** How does hyperparameter  $\lambda$  influence the performance of the proposed model?

### 5.1 Experiment Setup

**Downstream Task and Datasets.** We adopt the node classification task and evaluate the model on five widely used benchmark datasets: Cora, Citeseer, PubMed, Cornell, and Wisconsin. These datasets can be categorized into two distinct types based on their structural characteristics. In the first three datasets, nodes represent research publications, and edges denote citation relationships. In the remaining datasets, nodes correspond to webpages, and edges indicate hyperlinks between them. Unlike citation networks, webpage networks often exhibit different topological properties, making them valuable for testing the generalization ability of our approach. To simulate real-world adversarial scenarios, we implement MEAs on all five datasets during our experiments.

**GMLaaS Models.** We train a two-layer GCN as the target model within a GMLaaS setting. The model configuration is as follows: The hidden layer is configured with 16 features with ReLU activation, while the output layer uses softmax activation for classification. We optimize the model using the Adam optimizer with a learning rate of 0.01, a weight decay of 0.0005, and 200 training epochs. Following the transductive setting, the graph used during training is identical to the one used for inference.

**Adversarial Knowledge.** As previously discussed, we assume the attacker has partial knowledge of both the node attributes and graph structure. The adversary is allowed to access a single node and its one-hop subgraph at a time. Under this constraint, we implement three attack algorithms based on AL learning: AGE, GRAIN, and IGP, ensuring a realistic evaluation of our model's robustness against graph-based attacks.

**Baselines.** To thoroughly assess the effectiveness of the proposed detection framework, we compare it against a diverse set of baseline models, categorized as follows. We first employ commonly used neural network architectures for sequential and classification tasks: *Simple MLP* [32]: A fundamental feedforward neural network for classification. *RNN* [11]: A recurrent model that processes sequential data but struggles with long-term dependencies. *LSTM* [10]: An improved recurrent architecture incorporating gating mechanisms for long-range information retention. *Transformer* [39]: A self-attention-based architecture that effectively captures long-range dependencies in sequential data. Since MEAs detection can be framed as a time-series classification task [12], we incorporate several models from this domain: *Crossformer* [41]: Employs a cross-scale attention mechanism to capture temporal dependencies at different scales. *Autoformer* [46]: Integrates self-correlation and autoregressive structures to model periodic trends in time series. *TimesNet* [45]: Reformulates time series as a multi-period representation, capturing temporal variations both within and across periods. *PatchTST* [28]: Treats time-series segments as receptive fields in a convolutional framework, extracting multi-scale temporal patterns. *Informer* [53]: Uses a sparse attention mechanism to efficiently model long-range dependencies in sequential data.



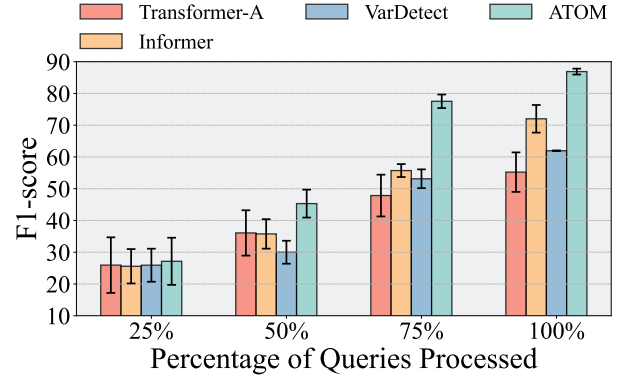
*iTransformer* [24]: A Transformer-based model that balances global temporal dependencies and local feature extraction through sequence decomposition. We also compare our framework with existing DNN-based detection approaches designed specifically for MEA detection in GMLaaS environments: *PRADA* [14]: Detects adversarial behavior by analyzing statistical deviations in sequential API queries. *VarDetect* [29]: Uses a Variational Autoencoder (VAE) to model user query distributions and identify anomalies.

**Ablated Models.** To assess the contribution of each individual component in the proposed framework, we conduct ablation studies by selectively removing or modifying specific modules. We define three ablated model variants: (1) *Replacing the enhanced GRU with a standard GRU*. This ablation removes the fusion gate, allowing us to evaluate the importance of differential input mechanisms. (2) *Replacing the proposed  $k$ -core-based embeddings with simple mean embeddings*. This experiment highlights the effectiveness of our scaling function, which is further explored in the Evaluation of Parameter Test section. (3) *Removing the mapping matrix*. This variation investigates the role of the mapping matrix in improving robustness by incorporating historical decision-making.

**Evaluation Metrics.** We evaluate the proposed framework using the following performance metrics: (1) *Detection Effectiveness*. We evaluate both the F1 score and recall metrics to measure attack detection performance. A higher F1 score and recall indicate better classification accuracy while minimizing false negatives. (2) *Ablation Study*. We compare the F1 scores of the full model and its ablated versions. This experiment also serves as an empirical validation of Theorem 4.4, particularly in analyzing the performance difference between the enhanced GRU and the standard GRU. (3) *Parameter Sensitivity*. We evaluate the impact of different values of the scaling factor  $\lambda$  on the F1 score. This analysis provides insights into how parameter tuning influences detection performance.

## 5.2 Evaluation of Detection Effectiveness

To address RQ1, we evaluate the effectiveness of ATOM by comparing its performance against multiple baselines. Since no existing methods are specifically designed for graph-based MEA detection in GMLaaS environments, we construct a diverse set of baselines to ensure a fair and comprehensive comparison. Specifically, we adopt classical classification models, replace the fusion GRU in ATOM with alternative sequential networks, and introduce time-series classification models to account for the temporal structure of the task. To maintain clarity, we append the suffix "-A" to the names of sequential baselines in Table 1. Additionally, we incorporate DNN-based detection strategies to examine the limitations of general MEA detection methods. To further evaluate real-time detection performance, we simulate progressive query arrival by assessing all models with {25%, 50%, 75%, 100%} of the query sequences. Here, we highlight the strongest-performing models within each category of the baselines. The performance of Transformer-A, Informer, VarDetect, and ATOM on Cora is visualized in Figure 2, while results for other baselines and datasets are provided in the Appendix. We summarize our observations below: (1) ATOM consistently achieves competitive performance across all baselines. It prioritizes recall value while maintaining a strong F1 score, which is particularly important for MEA detection. Specifically, ATOM achieves a



**Figure 2: Performance of Representative Models Over Sequential Query Processing on Cora.**

well-balanced distinction between attackers and legitimate users, enhancing its practical applicability. (2) DNN-based MEA detection methods do not generalize well to graph-based MEAs. In particular, PRADA exhibits the weakest performance among all models, as it assumes user queries follow a normal distribution, which is not realistic in real-world attack scenarios. Similarly, VarDetect, despite successfully encoding queries into a latent space, performs comparably to time-series classification models, underscoring the difficulty of directly extending existing DNN-based detection techniques to MEA detection in GNNs. (3) In real-time settings, ATOM outperforms all baselines at every percentage of queries processed while maintaining low variance. This highlights the advantage of processing queries sequentially rather than treating them as independent samples. During the first 25% of queries processed, all models exhibit similar performance due to the limited available information. However, as more queries are processed, the performance of ATOM rapidly improves, showcasing its ability to adapt dynamically to evolving attack strategies.

## 5.3 Evaluation of Ablation Study

To answer RQ2, we conduct a comprehensive evaluation of the full ATOM model and its ablated counterparts to assess the contribution of individual components. Specifically, we present their F1 scores across five benchmark datasets in Table 2. Through this analysis, we derive the following key observations: (1) The incorporation of second-order differences enhances MEA detection. This observation empirically supports Theorem 4.4, demonstrating that capturing higher-order temporal variations in user query sequences provides valuable discriminative features for identifying adversarial behavior. By modeling the second-order differences, the framework effectively captures subtle yet critical variations in query patterns, which would otherwise be overlooked by first-order representations. (2) The  $k$ -core embedding significantly improves performance compared to standard embedding. As shown in Table 2, the  $k$ -core embedding leads to a noticeable enhancement in classification, reinforcing its effectiveness in MEA detection. This improvement stems from its ability to extract topological features from the graph, which are particularly useful in distinguishing between queries. Furthermore, in the Evaluation of Parameter Test section, we provide a detailed discussion of how different levels of  $k$ -core embeddings

**Table 1: Performance Comparison between ATOM and baselines on different metrics and datasets. The best results are in bold.**

Metrics	F1 score					Recall				
	Wisconsin	Cornell	Cora	Citeseer	PubMed	Wisconsin	Cornell	Cora	Citeseer	PubMed
MLP	24.32 ± 15.72	56.71 ± 9.60	33.56 ± 13.74	39.06 ± 9.76	56.44 ± 7.93	23.57 ± 8.57	63.12 ± 14.32	34.15 ± 14.37	39.17 ± 10.34	58.42 ± 11.47
RNN-A	66.24 ± 8.83	52.08 ± 15.43	58.25 ± 4.70	59.05 ± 12.90	60.99 ± 10.92	53.57 ± 9.78	43.75 ± 15.93	52.36 ± 6.31	52.50 ± 17.50	51.92 ± 13.73
LSTM-A	53.12 ± 10.36	49.48 ± 10.03	57.58 ± 3.28	57.45 ± 10.89	54.24 ± 16.93	56.57 ± 9.70	44.53 ± 15.53	50.94 ± 3.53	49.17 ± 11.46	44.23 ± 18.34
Transformer-A	72.59 ± 7.05	60.24 ± 7.97	55.22 ± 6.22	70.10 ± 9.77	61.62 ± 10.28	60.71 ± 9.04	54.51 ± 9.09	49.37 ± 8.22	65.00 ± 14.34	51.58 ± 12.19
Crossformer	75.76 ± 0.24	67.79 ± 1.42	75.12 ± 4.35	59.69 ± 4.60	45.21 ± 17.12	79.29 ± 2.14	61.43 ± 16.2	63.87 ± 14.9	70.71 ± 15.2	46.09 ± 13.41
Autoformer	56.75 ± 7.67	79.07 ± 4.80	65.15 ± 17.5	60.76 ± 9.72	76.44 ± 10.8	55.71 ± 16.1	90.71 ± 9.61	57.42 ± 20.1	62.86 ± 19.6	75.22 ± 17.7
iTransformer	56.89 ± 6.90	55.98 ± 9.45	60.30 ± 3.12	62.08 ± 9.76	63.86 ± 3.97	60.00 ± 16.0	59.29 ± 17.3	51.77 ± 16.7	66.43 ± 14.8	63.04 ± 13.9
TimesNet	66.63 ± 5.90	81.79 ± 4.82	82.24 ± 6.79	59.22 ± 4.36	61.36 ± 0.18	68.57 ± 11.0	92.43 ± 6.59	84.52 ± 12.7	59.29 ± 13.4	53.91 ± 12.3
PatchTST	61.04 ± 8.40	62.96 ± 5.11	65.79 ± 12.9	57.56 ± 8.11	79.51 ± 5.39	62.86 ± 19.1	64.29 ± 16.5	59.84 ± 10.2	54.29 ± 12.6	82.61 ± 13.3
Informer	53.47 ± 0.42	81.36 ± 4.91	72.01 ± 1.63	49.24 ± 1.74	65.17 ± 4.37	54.29 ± 14.3	91.29 ± 4.14	71.45 ± 14.8	52.86 ± 17.1	63.04 ± 10.3
PRADA	19.01 ± 1.73	12.34 ± 0.89	11.23 ± 1.05	13.57 ± 1.52	16.78 ± 1.24	17.54 ± 1.42	13.45 ± 0.76	14.56 ± 0.98	15.89 ± 1.13	18.95 ± 1.37
VarDetect	64.28 ± 1.32	68.23 ± 24.7	61.95 ± 0.10	53.15 ± 1.24	55.16 ± 2.19	43.29 ± 1.14	60.74 ± 14.8	41.58 ± 0.71	52.17 ± 0.93	49.47 ± 2.75
ATOM	<b>81.48 ± 1.02</b>	<b>89.66 ± 0.97</b>	<b>86.88 ± 0.93</b>	<b>78.89 ± 1.39</b>	<b>83.24 ± 0.68</b>	<b>90.91 ± 3.67</b>	<b>96.15 ± 2.11</b>	<b>93.65 ± 1.07</b>	<b>85.71 ± 1.71</b>	<b>92.42 ± 2.02</b>

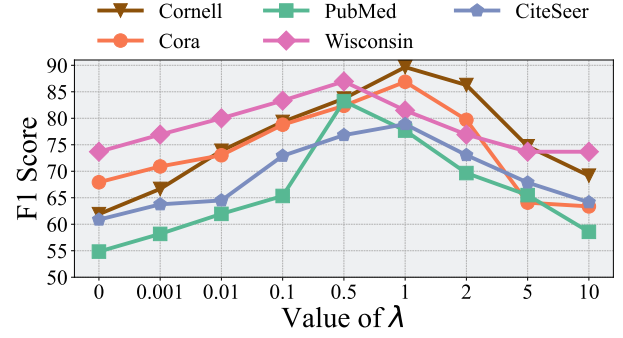
**Table 2: F1 scores from the ablated model on Cora, PubMed, and CiteSeer. The best results are highlighted in bold.**

Model	Cora	Citeseer	PubMed
ATOM	<b>86.88</b>	<b>78.89</b>	<b>83.24</b>
Standard GRU	80.64	71.97	75.47
Simple Embeddings	67.92	60.87	54.83
No Mapping Matrix	81.54	74.71	79.94

affect model performance, offering additional insights into the optimal choice of  $\lambda$ . (3) The mapping matrix acts as a specialized normalization mechanism for the hidden state, facilitating network convergence. More specifically, the mapping matrix functions as a scaling transformation applied to the hidden state, with its scaling factor dynamically controlled by the previous time step. This mechanism enhances the model's robustness by mitigating unstable fluctuations in hidden representations, thereby promoting stable and efficient convergence during training. The empirical results further confirm that the inclusion of the mapping matrix contributes to improved generalization performance.

#### 5.4 Evaluation of Parameter Test

To answer RQ3, we investigate the impact of varying the scaling factor  $\lambda$  on the proposed model across different datasets. Specifically, we systematically adjust  $\lambda$  in ATOM and report its F1 score in Figure 3. We record our results by  $\lambda \in \{0, 0.001, 0.01, 0.1, 0.5, 1, 2, 5, 10\}$ . The following key observations are drawn from this evaluation: (1) Incorporating  $\lambda$  as a scaling factor for the  $k$ -core value significantly enhances the ability of sequential networks to process queries. This result highlights the crucial role of topological information in refining node embeddings. By scaling the  $k$ -core value appropriately, ATOM effectively integrates structural properties into its feature representations, thereby improving its capacity to detect graph-based MEAs. The empirical results suggest that leveraging well-calibrated structural embeddings strengthens the temporal modeling capability of sequential networks, leading to more robust attack detection. (2) The selection of  $\lambda$  is critical; both excessively small and overly large values negatively impact model performance. When  $\lambda$  is too small, the influence of graph structural

**Figure 3: Impact of the adjustment factor  $\lambda$  in ATOM.**

information on node embeddings is minimal, making ATOM's performance comparable to that of models that solely rely on raw node attributes. This limitation prevents ATOM from fully exploiting the underlying network topology, thereby restricting its ability to capture adversarial patterns. When  $\lambda$  is set to a moderate value, ATOM achieves its best performance, with an improvement of up to 51.8% for PubMed in the F1 score compared to cases where no scaling factor is introduced ( $\lambda = 0$ ). In particular, when  $\lambda \approx 0.5$  to 1, ATOM effectively balances local attribute information with global topological properties, leading to more discriminative representations for MEA detection. This suggests that a well-calibrated  $\lambda$  allows the model to incorporate meaningful structural cues without overwhelming the influence of individual node features. When  $\lambda$  becomes excessively large, the F1 score exhibits a downward trend. This decline occurs because an overly strong emphasis on topological structure suppresses the contribution of node attribute features, leading to distorted representations. As a result, the model becomes less effective at distinguishing legitimate user queries from adversarial ones, ultimately impairing its detection capability. (3) The impact of  $\lambda$  varies across datasets, suggesting dataset-dependent optimal values. While an optimal range of  $\lambda \approx 0.5$  to 1 is generally observed, the exact value that maximizes performance may differ based on dataset-specific properties such as graph sparsity, node connectivity, and query distribution patterns. This indicates that tuning  $\lambda$  should be approached in a data-driven manner, potentially through cross-validation, to achieve the best trade-off between node attributes and topological information.



## 6 Conclusion

In this paper, we propose ATOM, a novel framework for detecting graph-based MEAs in GMLaaS environments under the transductive setting. To the best of our knowledge, we are the first to investigate the novel problem of detecting graph-based MEAs in GMLaaS. To address this problem, we design ATOM by focusing on real-time detecting and adaptive attacks. Specifically, we introduce sequential modeling and reinforcement learning to dynamically detect evolving attack patterns. We further conduct theoretical analysis for the query behavior and establish a theoretical foundation for our proposed framework. Extensive experiments on real-world datasets demonstrate ATOM's superior performance over baselines in the real-time detecting scenario. Meanwhile, two future directions are worth further investigation. First, we focus on the simulated queries in this paper due to the limited availability of query datasets in GMLaaS. Thus, exploring large-scale industrial query datasets from real-world scenarios could provide a more accurate reflection of practical attack behaviors. Second, to better capture realistic user behavior, it is essential to investigate distributed adversaries who coordinate attacks across multiple accounts, which could provide valuable insights for enhancing detection mechanisms.

## References

- [1] Pietro Bongini, Niccolò Pancino, Franco Scarselli, and Monica Bianchini. 2022. BioGNN: how graph neural networks can solve biological problems. In *Artificial Intelligence and Machine Learning for Healthcare: Vol. 1: Image and Data Analytics*. Springer, 211–231.
- [2] Hongyun Cai, Vincent W Zheng, and Kevin Chen-Chuan Chang. 2017. Active learning for graph embedding. *arXiv preprint arXiv:1705.05085* (2017).
- [3] Abhishek Chakraborty, Daniel Xing, Yuntao Liu, and Ankur Srivastava. 2022. Dynamarks: Defending against deep learning model extraction using dynamic watermarking. *arXiv preprint arXiv:2207.13321* (2022).
- [4] Chao Chang, Junming Zhou, Yu Weng, Xiangwei Zeng, Zhengyang Wu, Chang-Dong Wang, and Yong Tang. 2023. KGTN: Knowledge Graph Transformer Network for explainable multi-category item recommendation. *Knowledge-Based Systems* 278 (2023), 110854.
- [5] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014).
- [6] João Correia, João Capela, and Miguel Rocha. 2024. Deepmol: an automated machine and deep learning framework for computational chemistry. *Journal of Cheminformatics* 16, 1 (2024), 1–17.
- [7] Kamal A ElDahshan, Gaber E Abutaleb, Beriham R Elemery, Ebeid A Ebeid, and AbdAllah A AlHabshy. 2024. An optimized intelligent open-source MLaaS framework for user-friendly clustering and anomaly detection. *The Journal of Supercomputing* 80, 18 (2024), 26658–26684.
- [8] Chen Gao, Xiang Wang, Xiangnan He, and Yong Li. 2022. Graph neural networks for recommender system. In *Proceedings of the fifteenth ACM international conference on web search and data mining*. 1623–1625.
- [9] Faqian Guan, Tianqing Zhu, Hanjin Tong, and Wanlei Zhou. 2024. A realistic model extraction attack against graph neural networks. *Knowledge-Based Systems* 300 (2024), 112–144.
- [10] S Hochreiter. 1997. Long Short-term Memory. *Neural Computation MIT-Press* (1997).
- [11] John J Hopfield. 1982. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences* 79, 8 (1982), 2554–2558.
- [12] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. 2019. Deep learning for time series classification: a review. *Data mining and knowledge discovery* 33, 4 (2019), 917–963.
- [13] Hengrui Jia, Christopher A Choquette-Choo, Varun Chandrasekaran, and Nicolas Papernot. 2021. Entangled watermarks as a defense against model extraction. In *30th USENIX security symposium (USENIX Security 21)*. 1937–1954.
- [14] Mika Juuti, Sebastian Szlyler, Samuel Marchal, and N Asokan. 2019. PRADA: protecting against DNN model stealing attacks. In *2019 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, 512–527.
- [15] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
- [16] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).
- [17] Thomas N Kipf and Max Welling. 2016. Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308* (2016).
- [18] Isabell Lederer, Rudolf Mayer, and Andreas Rauber. 2023. Identifying appropriate intellectual property protection mechanisms for machine learning models: A systematization of watermarking, fingerprinting, model access, and attacks. *IEEE Transactions on Neural Networks and Learning Systems* (2023).
- [19] Taesung Lee, Benjamin Edwards, Ian Molloy, and Dong Su. 2019. Defending against neural network model stealing attacks using deceptive perturbations. In *2019 IEEE Security and Privacy Workshops (SPW)*. 43–49.
- [20] Jiacheng Liang, Ren Pang, Changjiang Li, and Ting Wang. 2024. Model extraction attacks revisited. In *Proceedings of the 19th ACM Asia Conference on Computer and Communications Security*. 1231–1245.
- [21] GuanJun Liu, Jing Tang, Yue Tian, and Jiacun Wang. 2021. Graph neural network for credit card fraud detection. In *2021 International Conference on Cyber-Physical Social Intelligence (ICCSII)*. IEEE, 1–6.
- [22] Xinjing Liu, Zhuo Ma, Yang Liu, Zhan Qin, Junwei Zhang, and Zhuzhu Wang. 2022. Selspect: Defending model stealing via heterogeneous semantic inspection. In *European Symposium on Research in Computer Security*. Springer, 610–630.
- [23] Yang Liu, Xiang Ao, Zidi Qin, Jianfeng Chi, Jinghua Feng, Hao Yang, and Qing He. 2021. Pick and choose: a GNN-based imbalanced learning approach for fraud detection. In *Proceedings of the web conference 2021*. 3168–3177.
- [24] Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. 2023. itransformer: Inverted transformers are effective for time series forecasting. *arXiv preprint arXiv:2310.06625* (2023).
- [25] Nils Lukas, Yuxuan Zhang, and Florian Kerschbaum. 2019. Deep neural network fingerprinting by conferrable adversarial examples. *arXiv preprint arXiv:1912.00888* (2019).
- [26] Pratyush Maini, Mohammad Yaghini, and Nicolas Papernot. 2021. Dataset inference: Ownership resolution in machine learning. *arXiv preprint arXiv:2104.10706* (2021).
- [27] Xuting Mao, Mingxi Liu, and Yinghui Wang. 2022. Using GNN to detect financial fraud based on the related party transactions network. *Procedia Computer Science* 124 (2022), 351–358.
- [28] Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. 2022. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730* (2022).
- [29] Soham Pal, Yash Gupta, Aditya Kanade, and Shirish Shevade. 2021. Stateful detection of model extraction attacks. *arXiv preprint arXiv:2107.05166* (2021).
- [30] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch. (2017).
- [31] Manon Réau, Nicolas Renaud, Li C Xue, and Alexandre MJJ Bonvin. 2023. DeepRank-GNN: a graph neural network framework to learn patterns in protein-protein interfaces. *Bioinformatics* 39, 1 (2023), btac759.
- [32] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. 1986. Learning representations by back-propagating errors. *nature* 323, 6088 (1986), 533–536.
- [33] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. 2008. The graph neural network model. *IEEE transactions on neural networks* 20, 1 (2008), 61–80.
- [34] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017).
- [35] Burr Settles. 2009. Active learning literature survey. (2009).
- [36] Jonathan Shlomi, Peter Battaglia, and Jean-Roch Vlimant. 2020. Graph neural networks in particle physics. *Machine Learning: Science and Technology* 2, 2 (2020), 021001.
- [37] Minxue Tang, Anna Dai, Louis DiValentin, Aolin Ding, Amin Hass, Neil Zhen-qiang Gong, and Yiran Chen. 2024. Modelguard: Information-theoretic defense against model extraction attacks. In *33rd USENIX Security Symposium (Security 2024)*.
- [38] Florian Tramèr, Fan Zhang, Ari Juels, Michael K Reiter, and Thomas Ristenpart. 2016. Stealing machine learning models via prediction {APIs}. In *25th USENIX security symposium (USENIX Security 16)*. 601–618.
- [39] A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* (2017).
- [40] Asim Waheed, Vasisht Duddu, and N. Asokan. 2024. GrOVe: Ownership Verification of Graph Neural Networks using Embeddings. In *2024 IEEE Symposium on Security and Privacy (SP)*. 2460–2477.
- [41] W Wang, L Yao, L Chen, B Lin, D Cai, X He, and W Liu. 2021. CrossFormer: A versatile vision transformer hinging on cross-scale attention. *arXiv 2021. arXiv preprint arXiv:2108.00154* (2021).
- [42] Qizhen Weng, Wencong Xiao, Yinghao Yu, Wei Wang, Cheng Wang, Jian He, Yong Li, Liping Zhang, Wei Lin, and Yu Ding. 2022. {MLaaS} in the wild: Workload analysis and scheduling in {Large-Scale} heterogeneous {GPU} clusters. In *19th USENIX Symposium on Networked Systems Design and Implementation (NSDI 22)*. 945–960.

- [43] Bang Wu, Xiangwen Yang, Shirui Pan, and Xingliang Yuan. 2022. Model extraction attacks on graph neural networks: Taxonomy and realisation. In *Proceedings of the 2022 ACM on Asia conference on computer and communications security*. 337–350.
- [44] Bang Wu, Xingliang Yuan, Shuo Wang, Qi Li, Minhui Xue, and Shirui Pan. 2024. Securing graph neural networks in mlaas: A comprehensive realization of query-based integrity verification. In *2024 IEEE Symposium on Security and Privacy (SP)*. 2534–2552.
- [45] Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. 2022. Timesnet: Temporal 2d-variation modeling for general time series analysis. *arXiv preprint arXiv:2210.02186* (2022).
- [46] Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. 2021. Autoformer: De-composition transformers with auto-correlation for long-term series forecasting. *Advances in neural information processing systems* 34 (2021), 22419–22430.
- [47] Sheng Xiang, Mingzhi Zhu, Dawei Cheng, Enxia Li, Ruihui Zhao, Yi Ouyang, Ling Chen, and Yefeng Zheng. 2023. Semi-supervised credit card fraud detection via attribute-driven graph representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 14557–14565.
- [48] Honggang Yu, Kaichen Yang, Teng Zhang, Yun-Yun Tsai, Tsung-Yi Ho, and Yier Jin. 2020. CloudLeak: Large-Scale Deep Learning Models Stealing Through Adversarial Examples. In *NDSS*, Vol. 38. 102.
- [49] Wentao Zhang, Yexin Wang, Zhenbang You, Meng Cao, Ping Huang, Jiulong Shan, Zhi Yang, and Bin Cui. 2022. Information gain propagation: a new way to graph active learning with soft labels. *arXiv preprint arXiv:2203.01093* (2022).
- [50] Wentao Zhang, Zhi Yang, Yexin Wang, Yu Shen, Yang Li, Liang Wang, and Bin Cui. 2021. Grain: Improving data efficiency of graph neural networks via diversified influence maximization. *arXiv preprint arXiv:2108.00219* (2021).
- [51] Xiao-Meng Zhang, Li Liang, Lin Liu, and Ming-Jing Tang. 2021. Graph neural networks and their current applications in bioinformatics. *Frontiers in genetics* 12 (2021), 690049.
- [52] Zhanyuan Zhang, Yizheng Chen, and David Wagner. 2021. SEAT: Similarity encoder by adversarial training for detecting model extraction attack queries. In *Proceedings of the 14th ACM Workshop on artificial intelligence and security*. 37–48.
- [53] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. 2021. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 11106–11115.

## A Proofs

**THEOREM A.1.** Consider a covering graph  $\mathcal{D}$  in the graph  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ , aiming to cover at least  $\beta \in [0, 1]$  percent nodes of  $\mathcal{G}$ , while minimizing  $\sum_{u \in \mathcal{A}} w(u)$ , where  $w(u)$  is the weight of node  $u$  and  $\mathcal{A}$  represents the set of nodes not being covered, then the maximum covering percentage is given by

$$\beta \leq \min\left\{\frac{|\mathcal{D}| - \frac{W}{\bar{w}_{\mathcal{A}}}}{\frac{n}{\delta} - \frac{W}{\bar{w}_{\mathcal{A}}}}, \frac{|\mathcal{D}| \cdot \delta}{n}\right\}. \quad (22)$$

Here,  $|\mathcal{D}|$  represents the number of nodes in  $\mathcal{D}$ ,  $W$  is the total weight in  $\mathcal{G}$ ,  $\bar{w}_{\mathcal{A}}$  represents the average weight in  $\mathcal{A}$  and  $\delta$  is the smallest degree for nodes in  $\mathcal{D}$ .

**PROOF.** Since  $\mathcal{D}$  at least covers  $\beta$  of  $\mathcal{G}$ , we then get

$$|\mathcal{D}| \cdot \delta \geq \beta \cdot n. \quad (23)$$

Also,

$$|\mathcal{D}| \leq \frac{\beta \cdot n}{\delta} + \frac{W_{\mathcal{A}}}{\bar{w}_{\mathcal{A}}}, \quad (24)$$

where  $W_{\mathcal{A}} \leq (1 - \beta) \cdot n \cdot \bar{w}_{\mathcal{V}}$  is defined as the total weight of  $\mathcal{A}$ , as  $\bar{w}_{\mathcal{V}}$  representing the total average weight in  $\mathcal{V}$ ,  $\bar{w}_{\mathcal{A}}$  is the average weight in  $\mathcal{A}$ . By (23) and (24), we directly get

$$\beta \leq \frac{|\mathcal{D}| \cdot \delta}{n}. \quad (25)$$

By solving (25) and  $W_{\mathcal{A}}$  simultaneously, we obtain:

$$|\mathcal{D}| - n \frac{\bar{w}_{\mathcal{V}}}{\bar{w}_{\mathcal{A}}} \leq n\beta\left(\frac{1}{\delta} - \frac{\bar{w}_{\mathcal{V}}}{\bar{w}_{\mathcal{A}}}\right) \quad (26)$$

Then if  $\frac{\bar{w}_{\mathcal{V}}}{\bar{w}_{\mathcal{A}}} > \frac{1}{\delta}$ ,

$$\beta \leq \frac{\frac{|\mathcal{D}| \cdot \delta}{n} - \delta \frac{\bar{w}_{\mathcal{V}}}{\bar{w}_{\mathcal{A}}}}{1 - \delta \frac{\bar{w}_{\mathcal{V}}}{\bar{w}_{\mathcal{A}}}}, \quad (27)$$

otherwise, i.e.,  $0 < \frac{\bar{w}_{\mathcal{V}}}{\bar{w}_{\mathcal{A}}} < \frac{1}{\delta}$ , we have

$$\beta \geq \frac{\frac{|\mathcal{D}| \cdot \delta}{n} - \delta \frac{\bar{w}_{\mathcal{V}}}{\bar{w}_{\mathcal{A}}}}{1 - \delta \frac{\bar{w}_{\mathcal{V}}}{\bar{w}_{\mathcal{A}}}}. \quad (28)$$

Thus, by (25), (27) and (28),

$$\beta \leq \min\left\{\frac{\frac{|\mathcal{D}|}{n} - \frac{\bar{w}_{\mathcal{V}}}{\bar{w}_{\mathcal{A}}}}{\frac{1}{\delta} - \frac{\bar{w}_{\mathcal{V}}}{\bar{w}_{\mathcal{A}}}}, \frac{|\mathcal{D}| \cdot \delta}{n}\right\}. \quad (29)$$

Introducing  $\frac{\bar{w}_{\mathcal{V}}}{\bar{w}_{\mathcal{A}}} = \frac{\bar{w}_{\mathcal{V}} n}{\bar{w}_{\mathcal{A}} n} = \frac{W}{\bar{w}_{\mathcal{A}} n}$ , we finally finish the proof

$$\beta \leq \min\left\{\frac{\frac{|\mathcal{D}|}{n} - \frac{W}{\bar{w}_{\mathcal{A}} n}}{\frac{1}{\delta} - \frac{W}{\bar{w}_{\mathcal{A}} n}}, \frac{|\mathcal{D}| \cdot \delta}{n}\right\}. \quad (30)$$

□

**PROPOSITION A.2.** Consider a changing graph  $\mathcal{D}_{t-1}$  and  $\mathcal{D}_t$  in the graph  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ , where  $\mathcal{D}_{t-1} \subset \mathcal{D}_t \subset \mathcal{G}$ , achieving at least  $\beta_{t-1}$  covering rate with the lowest degree  $\delta_{t-1}$  and at most  $\beta_t$  covering rate with the lowest degree  $\delta_t$ , respectively. Also, suppose that  $\mathcal{A}_{t-1}$  and  $\mathcal{A}_t$  represent the set of nodes that are not covered, with the corresponding weight  $W_{\mathcal{A}_{t-1}}$ ,  $W_{\mathcal{A}_t}$  and the average weight  $\bar{w}_{\mathcal{A}_{t-1}}$ ,  $\bar{w}_{\mathcal{A}_t}$ . We then get

$$\frac{\Delta W_{\mathcal{A}}}{\Delta |\mathcal{D}|} \leq \frac{(\beta_t - \beta_{t-1}) \cdot W}{|\mathcal{D}_{t-1}|(1 - \frac{\delta_{t-1}}{\delta_t})}, \quad (31)$$

and  $\delta_t > \delta_{t-1}$ . Here,  $|\mathcal{D}_{t-1}|$ ,  $|\mathcal{D}_t|$  represents the number of nodes in  $\mathcal{D}_{t-1}$ ,  $\mathcal{D}_t$ ,  $W$  is the total weight in  $\mathcal{G}$ .

**PROOF.** Still, we directly have

$$\begin{cases} |\mathcal{D}_{t-1}| \cdot \delta_{t-1} \geq \beta_{t-1} \cdot n, \\ |\mathcal{D}_t| \cdot \delta_t \leq \beta_t \cdot n, \end{cases} \quad (32)$$

and

$$\begin{cases} W_{\mathcal{A}_{t-1}} \leq (1 - \beta_{t-1}) \cdot n \cdot \bar{w}_{\mathcal{V}}, \\ W_{\mathcal{A}_t} \geq (1 - \beta_t) \cdot n \cdot \bar{w}_{\mathcal{V}}. \end{cases} \quad (33)$$

By solving (33), we obtain:

$$0 < W_{\mathcal{A}_{t-1}} - W_{\mathcal{A}_t} \leq (\beta_t - \beta_{t-1}) \cdot n \cdot \bar{w}_{\mathcal{V}}. \quad (34)$$

From (32) and assuming  $|\mathcal{D}_t| \cdot \delta_t \geq |\mathcal{D}_{t-1}| \cdot \delta_{t-1}$ , we have

$$|\mathcal{D}_{t-1}|(1 - \frac{\delta_{t-1}}{\delta_t}) \leq |\mathcal{D}_t| - |\mathcal{D}_{t-1}| \leq n \cdot (\frac{\beta_t}{\delta_t} - \frac{\beta_{t-1}}{\delta_{t-1}}) \quad (35)$$

By (34) and (35), we get

$$\frac{W_{\mathcal{A}_{t-1}} - W_{\mathcal{A}_t}}{|\mathcal{D}_t| - |\mathcal{D}_{t-1}|} \leq \frac{(\beta_t - \beta_{t-1}) \cdot W}{|\mathcal{D}_{t-1}|(1 - \frac{\delta_{t-1}}{\delta_t})}, \quad (36)$$

that is,

$$\frac{\Delta W_{\mathcal{A}}}{\Delta |\mathcal{D}|} \leq \frac{(\beta_t - \beta_{t-1}) \cdot W}{|\mathcal{D}_{t-1}|(1 - \frac{\delta_{t-1}}{\delta_t})}, \quad (37)$$

which requires that

$$\delta_t > \delta_{t-1}. \quad (38)$$

□

PROPOSITION A.3. Consider changing graphs  $\mathcal{D}_{t-1}$ ,  $\mathcal{D}_t$  and  $\mathcal{D}_{t+1}$  in the graph  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ , where  $\mathcal{D}_{t-1} \subset \mathcal{D}_t \subset \mathcal{D}_{t+1} \subset \mathcal{G}$ , achieving at least  $\beta_{t-1}$  covering rate with the lowest degree  $\delta_{t-1}$  and at most  $\beta_t$  covering rate with the lowest degree  $\delta_t$ . Also, there is at least  $\beta'_t$  covering rate and at most  $\beta_{t+1}$  covering rate at time step  $t$  and  $t+1$ . Suppose that  $\mathcal{A}_{t-1}$ ,  $\mathcal{A}_t$  and  $\mathcal{A}_{t+1}$  represent the set of nodes that are not covered, with the corresponding weight  $W_{\mathcal{A}_{t-1}}$ ,  $W_{\mathcal{A}_t}$ ,  $W_{\mathcal{A}_{t+1}}$  and the average weight  $w_{\mathcal{A}_{t-1}}$ ,  $w_{\mathcal{A}_t}$ ,  $w_{\mathcal{A}_{t+1}}$ . We then get

$$\frac{\Delta^2 W_{\mathcal{A}}}{\Delta^2 |\mathcal{D}|} \geq \left| \frac{W_d}{n} \frac{\Delta \delta_t}{\Delta \beta_t} - \frac{W_{\delta_{t+1}}}{|\mathcal{D}_t|} \frac{\Delta \beta_{t+1}}{\Delta \delta_{t+1}} \right|, \quad (39)$$

and  $\delta_{t+1} > \delta_t > \delta_{t-1}$ . Here,  $|\mathcal{D}_{t-1}|$ ,  $|\mathcal{D}_t|$  and  $|\mathcal{D}_{t+1}|$  represent the number of nodes in  $\mathcal{D}_{t-1}$ ,  $\mathcal{D}_t$  and  $\mathcal{D}_{t+1}$ ,  $W$  is the total weight in  $\mathcal{G}$ , assuming  $W_{\mathcal{A}_{t-1}} - W_{\mathcal{A}_t} \geq W_d$ .

PROOF. By the discussion from Proposition A.2, we have

$$\frac{W_{\mathcal{A}_t} - W_{\mathcal{A}_{t+1}}}{|\mathcal{D}_{t+1}| - |\mathcal{D}_t|} \leq \frac{(\beta_{t+1} - \beta'_t) \cdot W}{|\mathcal{D}_t| (1 - \frac{\delta_t}{\delta_{t+1}})}. \quad (40)$$

Assuming that we have known  $W_{\mathcal{A}_{t-1}} - W_{\mathcal{A}_t} \geq W_d$ , then we get

$$\frac{W_{\mathcal{A}_{t-1}} - W_{\mathcal{A}_t}}{|\mathcal{D}_t| - |\mathcal{D}_{t-1}|} \geq \frac{W_d}{n \cdot (\frac{\beta_t}{\delta_t} - \frac{\beta_{t-1}}{\delta_{t-1}})}. \quad (41)$$

The difference between two inequality is given by

$$\frac{\Delta^2 W_{\mathcal{A}}}{\Delta^2 |\mathcal{D}|} = \left| \frac{W_{\mathcal{A}_{t-1}} - W_{\mathcal{A}_t}}{|\mathcal{D}_t| - |\mathcal{D}_{t-1}|} - \frac{W_{\mathcal{A}_t} - W_{\mathcal{A}_{t+1}}}{|\mathcal{D}_{t+1}| - |\mathcal{D}_t|} \right|, \quad (42)$$

which is restrict by

$$\frac{\Delta^2 W_{\mathcal{A}}}{\Delta^2 |\mathcal{D}|} \geq \left| \frac{W_d}{n \cdot (\frac{\beta_t}{\delta_t} - \frac{\beta_{t-1}}{\delta_{t-1}})} - \frac{(\beta_{t+1} - \beta'_t) \cdot W}{|\mathcal{D}_t| (1 - \frac{\delta_t}{\delta_{t+1}})} \right|. \quad (43)$$

Observe that  $\beta_t \frac{\delta_{t-1}}{\delta_t} + \beta_{t-1} \frac{\delta_t}{\delta_{t-1}} > 2\beta_{t-1}$ , then

$$\frac{\Delta^2 W_{\mathcal{A}}}{\Delta^2 |\mathcal{D}|} \geq \left| \frac{W_d (\delta_t - \delta_{t-1})}{n (\beta_t - \beta_{t-1})} - \frac{W_{\delta_{t+1}} (\beta_{t+1} - \beta'_t)}{|\mathcal{D}_t| (\delta_{t+1} - \delta_t)} \right|. \quad (44)$$

Introduce  $\Delta \beta_t = \beta_t - \beta_{t-1}$  and  $\Delta \delta_t = \delta_t - \delta_{t-1}$ . We get

$$\frac{\Delta^2 W_{\mathcal{A}}}{\Delta^2 |\mathcal{D}|} \geq \left| \frac{W_d \Delta \delta_t}{n \Delta \beta_t} - \frac{W_{\delta_{t+1}} \Delta \beta_{t+1}}{|\mathcal{D}_t| \Delta \delta_{t+1}} \right|. \quad (45)$$

By Triangle Inequality, we finally have

$$\frac{\Delta^2 W_{\mathcal{A}}}{\Delta^2 |\mathcal{D}|} \geq \left| \frac{W_d}{n} \frac{\Delta \delta_t}{\Delta \beta_t} - \frac{W_{\delta_{t+1}}}{|\mathcal{D}_t|} \frac{\Delta \beta_{t+1}}{\Delta \delta_{t+1}} \right|. \quad (46)$$

□

THEOREM A.4. Consider changing graphs  $\mathcal{D}_{t-1}$ ,  $\mathcal{D}_t$  and  $\mathcal{D}_{t+1}$  with graph  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ , where  $\mathcal{D}_{t-1} \subset \mathcal{D}_t \subset \mathcal{D}_{t+1} \subset \mathcal{G}$ , aiming at achieving at least  $\beta_{t-1}$  covering rate with the lowest degree  $\delta_{t-1}$  and at most  $\beta_t$  covering rate with the lowest degree  $\delta_t$ . Also, there is at least  $\beta'_t$  covering rate and at most  $\beta_{t+1}$  covering rate at time step  $t$  and  $t+1$ . Suppose that  $\mathcal{A}_{t-1}$ ,  $\mathcal{A}_t$  and  $\mathcal{A}_{t+1}$  represent the set of nodes that are not covered, with the corresponding weight  $W_{\mathcal{A}_{t-1}}$ ,  $W_{\mathcal{A}_t}$ ,  $W_{\mathcal{A}_{t+1}}$  and the average weight  $w_{\mathcal{A}_{t-1}}$ ,  $w_{\mathcal{A}_t}$ ,  $w_{\mathcal{A}_{t+1}}$ . The second-order differences become essential, that is,  $\frac{\Delta^2 W_{\mathcal{A}}}{\Delta^2 |\mathcal{D}|} \geq \frac{\Delta W_{\mathcal{A}}}{\Delta |\mathcal{D}|}$  holds when

$$W_d \geq \frac{n \delta_{t+1} \Delta \beta_t}{|\mathcal{D}_{t-1}| \Delta \delta_t} \left( \frac{\Delta \beta_t}{\Delta \delta_t} + \frac{\Delta \beta_{t+1}}{\Delta \delta_{t+1}} \right) W, \quad (47)$$

Table 3: Statistics of the adopted real-world graph datasets.

	#Nodes	#Edges	#Attributes	#Classes
Cora	2,708	5,429	1,433	7
CiteSeer	3,327	4,723	3,703	6
PubMed	19,717	88,648	500	3
Cornell	183	298	1703	5
Wisconsin	251	515	1703	5

and  $\delta_{t+1} > \delta_t > \delta_{t-1}$ . Here,  $|\mathcal{D}_{t-1}|$ ,  $|\mathcal{D}_t|$  and  $|\mathcal{D}_{t+1}|$  represent the number of nodes in  $\mathcal{D}_{t-1}$ ,  $\mathcal{D}_t$  and  $\mathcal{D}_{t+1}$ ,  $W$  is the total weight in  $\mathcal{G}$ , assuming  $W_{\mathcal{A}_{t-1}} - W_{\mathcal{A}_t} \geq W_d$ .

PROOF. By proposition A.2 and A.3, we write out that

$$\frac{(\beta_t - \beta_{t-1}) \cdot W}{|\mathcal{D}_{t-1}| (1 - \frac{\delta_{t-1}}{\delta_t})} \leq \left| \frac{W_d}{n} \frac{\Delta \delta_t}{\Delta \beta_t} - \frac{W_{\delta_{t+1}}}{|\mathcal{D}_t|} \frac{\Delta \beta_{t+1}}{\Delta \delta_{t+1}} \right|, \quad (48)$$

which gives

$$W_d \geq \frac{n \Delta \beta_t W}{\Delta \delta_t} \left| \frac{\delta_t}{|\mathcal{D}_{t-1}|} \frac{\Delta \beta_t}{\Delta \delta_t} + \frac{\delta_{t+1}}{|\mathcal{D}_t|} \frac{\Delta \beta_{t+1}}{\Delta \delta_{t+1}} \right|. \quad (49)$$

By Triangle Inequality,  $W_d$  is required by

$$W_d \geq \frac{n \Delta \beta_t W}{\Delta \delta_t} \left( \frac{\delta_t}{|\mathcal{D}_{t-1}|} \frac{\Delta \beta_t}{\Delta \delta_t} + \frac{\delta_{t+1}}{|\mathcal{D}_t|} \frac{\Delta \beta_{t+1}}{\Delta \delta_{t+1}} \right). \quad (50)$$

For simplicity, we can further get a tighter bound required by

$$W_d \geq \frac{n \delta_{t+1} \Delta \beta_t}{|\mathcal{D}_{t-1}| \Delta \delta_t} \left( \frac{\Delta \beta_t}{\Delta \delta_t} + \frac{\Delta \beta_{t+1}}{\Delta \delta_{t+1}} \right) W. \quad (51)$$

□

## B Reproducibility

This section provides detailed descriptions of our datasets, experimental settings, and implementation details to ensure the reproducibility of our experiments. The full implementation, including code and configuration files, is available in our repository <https://github.com/LabRAI/ATOM>.

### B.1 Real-World Datasets.

We conduct experiments using multiple widely adopted node classification datasets. The key statistics of these datasets are summarized in Table 3.

### B.2 Experimental Settings.

For each real-world dataset in our experiment, we adopt Active-Learning-based MEAs to generate query sequences for the detection task. Here we set the hyperparameters in AL-based MEAs to be a wide list of values, where we present varying values of {1%, 5%, 10%, 15%} percentage of the nodes in the graph as a prior knowledge and {35, 70, 105, 140, 200, 300, 400, 500} query budgets. We note that query budgets are always smaller than the nodes in the subgraph, and different sizes of the dataset will allow different numbers of query budgets. While the query sequence is generated, the fidelity of the extracted model is given to help label the sequence it is from. Generally, we label the sequence as an attacker if it corresponds to a fidelity larger than 0.65 and a long query sequence, otherwise, we label it as a legitimate user if the query

sequence is short or there is a fidelity smaller than 0.2. We split all the query sequences from the same dataset with 70% for training, 15% for validating, and 15% for testing. Only the sequence labels in the training set are visible for all models during training. For different datasets, the hyperparameters vary, but keep the same for the proposed model and its baselines. For all datasets in our experiment, we train a two-layer GCN by 200 epochs as our GMLaaS system. And we adopt a learning rate of 0.01 and a weight decay of 0.0005 while training.

### B.3 Implementation of ATOM.

ATOM is implemented based on Pytorch [30] with Adam optimizer [15]. To ensure a fair and comprehensive evaluation, we conduct experiments using multiple random seeds and analyze model performance under different initialization conditions. Additionally, we perform an extensive hyperparameter search over ATOM's parameter space, including the learning rate (lr), hidden state dimension, PPO clipping parameter, entropy coefficient, and lambda ( $\lambda$ ). For consistency, the same number of hyperparameter searches is conducted for all baseline methods, and we report the best-performing configuration along with the standard deviation for each method. To accelerate training, we utilize four NVIDIA RTX 4090 GPUs for synchronous training, which significantly reduces the training time. However, it is important to note that different hardware configurations may lead to variations in reproducibility.

### B.4 Implementation of Baselines.

**MLP.** We use a two-hidden-layer MLP for binary classification. Note that the MLP cannot process time series, we adopt the mean and the max features of the sequences instead.

**RNN-A, LSTM-A and Transformer-A.** We replace the sequential structure in ATOM with RNN, LSTM, and Transformer, named RNN-A, LSTM-A, and Transformer-A correspondingly, which thus allows us to classify the sequences. It should be mentioned that we adopt sine position coding and 4 multi-head attention in the implementation of the Transformer-A.

**Crossformer, Autoformer, iTransformer, TimeNet, PatchTST and Informer.** We adopt a series of well-established baseline models in the field of time-series forecasting and classification. And we adopt their official open-source code for experiments<sup>1</sup>.

**PRADA.** We adopt a statistical method as a baseline for comparison. We note that it is defined on static sequences, and we follow its official open-source code for experiments<sup>2</sup>.

**VarDetect.** We adopt an effective MEAs detection method based on Var as a baseline for comparison. We note that it is defined on static sequences and allows three types of outputs, specifically, VarDetect may output Alarm, Normal, and Uncertain. We follow its official open-source code for experiments<sup>3</sup>.

### B.5 Packages Required for Implementations.

We perform the experiments mainly on a server with multiple Nvidia 4090 GPUs. We list the main packages with their versions in our repository.

<sup>1</sup><https://github.com/thuml/Time-Series-Library>

<sup>2</sup><https://github.com/SSGAalto/prada-protecting-against-dnn-model-stealing-attacks>

<sup>3</sup><https://github.com/vardetect/vardetect>

## C Supplementary Experiments

### C.1 Evaluation of Detection Effectiveness

In this subsection, we provide additional experimental results regarding the real-time detection effectiveness on different models and datasets. Specifically, we have shown the performance evolution over sequential query processing on Cora in Section 5.2, and here we present more comprehensive results in Table 4, Table 5, and Table 6, with all other settings being consistent with the experiments presented in Section 5.2.

**Table 4: Detection Performance with 25% Query Sequences Across Different Datasets**

Metrics	F1 score				
	Wisconsin	Cornell	Cora	Citeseer	PubMed
MLP	16.67 ± 8.17	28.88 ± 8.72	13.73 ± 8.69	10.14 ± 9.85	12.44 ± 8.41
RNN-A	15.72 ± 7.41	18.17 ± 7.74	26.86 ± 9.01	29.54 ± 8.78	12.83 ± 7.42
LSTM-A	20.86 ± 7.16	23.64 ± 7.69	18.86 ± 8.34	21.26 ± 9.11	22.61 ± 7.06
Transformer-A	30.69 ± 7.82	22.34 ± 8.63	25.93 ± 7.74	23.39 ± 9.14	14.29 ± 7.21
Crossformer	13.72 ± 6.11	19.15 ± 7.62	20.37 ± 7.32	19.77 ± 8.26	17.64 ± 9.34
Autoformer	11.72 ± 7.45	12.34 ± 6.78	14.56 ± 8.90	16.78 ± 9.01	18.90 ± 8.23
iTransformer	13.45 ± 8.79	15.67 ± 7.89	17.89 ± 7.34	19.01 ± 8.45	21.23 ± 9.56
TimesNet	14.56 ± 7.01	16.78 ± 8.90	18.90 ± 7.45	21.23 ± 9.56	23.45 ± 10.67
PatchTST	15.67 ± 6.87	17.89 ± 9.12	20.12 ± 6.23	22.34 ± 7.34	24.56 ± 9.45
Informer	13.46 ± 9.12	15.58 ± 8.65	25.58 ± 8.39	20.55 ± 6.52	21.74 ± 8.32
PRADA	4.96 ± 2.72	6.25 ± 2.94	6.38 ± 3.56	5.52 ± 2.23	4.87 ± 1.41
VarDetect	17.23 ± 10.31	23.78 ± 10.74	25.91 ± 7.90	15.90 ± 9.31	29.65 ± 10.34
ATOM	34.91 ± 6.42	28.12 ± 6.83	27.14 ± 7.79	34.59 ± 6.41	25.47 ± 3.49

**Table 5: Detection Performance with 50% Query Sequences Across Different Datasets.**

Metrics	F1 score				
	Wisconsin	Cornell	Cora	Citeseer	PubMed
MLP	17.91 ± 8.57	34.44 ± 9.36	14.69 ± 12.84	23.73 ± 10.08	30.07 ± 10.69
RNN-A	29.71 ± 7.72	26.05 ± 8.44	33.74 ± 18.16	46.51 ± 9.16	29.89 ± 9.64
LSTM-A	34.42 ± 8.26	34.67 ± 8.11	34.16 ± 19.43	31.42 ± 8.85	31.42 ± 9.12
Transformer-A	54.58 ± 6.21	7.55 ± 13.23	36.07 ± 10.14	36.74 ± 9.29	32.26 ± 8.21
Crossformer	37.89 ± 8.95	35.77 ± 8.76	35.75 ± 8.85	34.61 ± 7.22	36.51 ± 8.38
Autoformer	34.44 ± 9.82	33.73 ± 8.73	23.71 ± 8.55	30.06 ± 7.24	29.72 ± 8.74
iTransformer	35.55 ± 7.90	38.25 ± 8.41	37.00 ± 9.00	36.15 ± 7.12	39.75 ± 6.50
TimesNet	32.10 ± 8.05	31.85 ± 7.25	31.78 ± 9.30	31.52 ± 6.50	34.35 ± 8.75
PatchTST	37.10 ± 8.10	35.50 ± 7.50	36.00 ± 8.70	34.85 ± 8.85	35.40 ± 6.42
Informer	37.91 ± 8.98	35.78 ± 8.78	35.76 ± 7.86	34.62 ± 6.23	46.52 ± 7.39
PRADA	9.17 ± 3.03	7.45 ± 3.52	8.20 ± 4.77	8.08 ± 4.01	7.95 ± 4.26
VarDetect	39.12 ± 11.44	33.34 ± 7.07	30.00 ± 11.76	36.79 ± 7.92	49.19 ± 11.91
ATOM	49.77 ± 6.65	49.18 ± 5.39	45.31 ± 6.94	40.25 ± 4.48	35.49 ± 7.87

**Table 6: Detection Performance with 75% Query Sequences Across Different Datasets.**

Metrics	F1 score				
	Wisconsin	Cornell	Cora	Citeseer	PubMed
MLP	20.61 ± 9.74	45.37 ± 9.44	18.03 ± 10.13	33.26 ± 13.41	43.82 ± 9.87
RNN-A	56.30 ± 9.21	43.82 ± 8.87	45.40 ± 8.18	50.01 ± 8.21	43.50 ± 8.54
LSTM-A	44.78 ± 9.84	37.75 ± 8.54	48.15 ± 9.06	47.62 ± 10.65	42.55 ± 8.23
Transformer-A	48.51 ± 8.86	49.76 ± 7.80	47.84 ± 7.78	41.14 ± 8.46	59.62 ± 7.97
Crossformer	47.75 ± 7.85	55.21 ± 8.98	52.69 ± 9.53	46.43 ± 9.34	41.76 ± 10.47
Autoformer	42.18 ± 7.25	63.47 ± 8.54	50.12 ± 8.01	45.98 ± 9.76	61.42 ± 8.89
iTransformer	41.52 ± 7.34	46.89 ± 9.15	54.37 ± 7.78	47.12 ± 8.98	51.23 ± 7.56
TimesNet	41.96 ± 8.12	72.43 ± 7.21	55.87 ± 8.87	45.74 ± 8.15	57.18 ± 8.63
PatchTST	45.34 ± 8.56	58.12 ± 8.45	61.23 ± 7.02	48.34 ± 7.34	59.12 ± 8.01
Informer	44.72 ± 7.98	67.77 ± 9.21	65.71 ± 8.12	44.07 ± 9.71	59.76 ± 7.19
PRADA	13.45 ± 1.56	10.04 ± 0.87	10.15 ± 1.24	10.16 ± 2.56	14.56 ± 1.12
VarDetect	49.92 ± 9.13	48.37 ± 9.72	53.13 ± 7.07	47.17 ± 7.43	50.79 ± 6.94
ATOM	74.13 ± 4.46	69.32 ± 5.39	77.52 ± 4.10	60.48 ± 5.65	71.91 ± 4.81

**Table 7: F1 scores from the ablated model. The best results are highlighted in bold.**

Model	Cornell	Cora	PubMed	Wisconsin	Citeseer
ATOM	<b>89.66</b>	<b>86.88</b>	<b>83.24</b>	<b>81.48</b>	<b>78.89</b>
Standard GRU	85.19	80.64	75.47	73.68	71.97
Simple Embeddings	61.90	67.92	54.83	73.41	60.87
No Mapping Matrix	61.31	81.54	79.94	75.93	74.71

## C.2 Evaluation of Ablation Study

In this subsection, we provide additional experimental results for the ablation study. Specifically, we have presented F1 scores from the ablated model on Cora, PubMed, and CiteSeer in Table 2. Here we show the other two datasets named Wisconsin and Cornell in Table 7 to show the generalization of our experiments.

## D Probabilistic Interpretation of ATOM

In this section, we provide a probabilistic interpretation of our model. In particular, we consider the query sequences generated by users and show how they can be organized and analyzed as a stochastic process.

**Definition D.1 (Query Lists).** For each user  $u_i$ , suppose the observed query sequence is  $Q_i = \{q_{i,1}, q_{i,2}, \dots, q_{i,T_i}\}$ . From each sequence  $Q_i$ , we construct a corresponding query list  $l_i$  by sequentially connecting consecutive queries. That is, for each  $p \in \{2, 3, \dots, T_i\}$  we connect  $q_{i,p-1}$  to  $q_{i,p}$  with an edge whose weight  $w_{i,p-1}$  is defined as the length of the shortest path from  $q_{i,p-1}$  to  $q_{i,p}$  on the graph  $\mathcal{G}$ . We denote the number of nodes in  $l_i$  by its length  $|l_i|$ .

**PROPOSITION D.2.** Consider the graph  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$  used for training in GMLaaS under the transductive setting. Assume that, due to the diversity of prior knowledge among normal users, every node in  $\mathcal{G}$  is eventually visited by some user. Then there exists a collection of query lists generated by normal users whose union covers  $\mathcal{V}$  and which are pairwise disjoint (i.e., no two lists share any node). In fact, if we denote by  $l_{\min}$  a query list having the minimum number of nodes, then by the pigeonhole principle the maximum number of pairwise disjoint query lists is bounded by  $J = \lceil \frac{|\mathcal{G}|}{l_{\min}} \rceil$ .

**PROOF.** Since all nodes in  $\mathcal{G}$  are visited by some normal user, we can extract query lists so that every node appears in at least one list. Choosing one list  $l_{\min}$  that is shortest (i.e., has the minimum number of nodes), note that any collection of pairwise disjoint query lists must assign at least  $|l_{\min}|$  distinct nodes to each list. Hence, by the pigeonhole principle the number of such disjoint lists is at most  $\lceil \frac{|\mathcal{G}|}{l_{\min}} \rceil$ .  $\square$

Let us now denote this upper bound by  $J = \lceil \frac{|\mathcal{G}|}{l_{\min}} \rceil$ . We select a collection of  $J$  query lists  $\{l_i\}_{i=1}^J$  from the normal users. To facilitate further analysis, we pad each query list so that every list has the same length. Specifically, let  $k = \max\{|l_1|, |l_2|, \dots, |l_J|\}$ . For any query list  $l_i = \{q_{i,1}, q_{i,2}, \dots, q_{i,T_i}\}$  with  $|l_i| < k$ , we extend it by replicating its last query  $q_{i,T_i}$  for positions  $T_i+1, T_i+2, \dots, k$  and assign an edge weight of 0 to each newly introduced edge. This padding ensures that each list  $l_i$  is represented as a sequence of exactly  $k$  queries. Observe that for the  $(J+1)$ th query list, every

query it contains already appears in one of the first  $J$  lists. Thus, we can regard the generation of query sequences as a stochastic process over the collection  $\{l_i\}_{i=1}^J$ . We now introduce several definitions that formalize this process.

**Definition D.3 (List Distance).** For any two padded query lists

$$l_i = \{q_{i,1}, q_{i,2}, \dots, q_{i,k}\} \quad l_j = \{q_{j,1}, q_{j,2}, \dots, q_{j,k}\}, \quad (52)$$

the distance between  $l_i$  and  $l_j$  is defined as

$$d(i, j) = \sum_{s=1}^k |q_{i,s} \rightarrow q_{j,s}|, \quad (53)$$

where  $|q_{i,s} \rightarrow q_{j,s}|$  denotes the length of the shortest path on  $\mathcal{G}$  between the  $s$ th query of  $l_i$  and the  $s$ th query of  $l_j$ .

**Definition D.4 (List Transition Probability).** Given the distance  $d(i, j)$  and a sensitivity parameter  $\lambda_s > 0$ , the probability of transitioning from query list  $l_i$  to query list  $l_j$  is defined according to the Boltzmann distribution as

$$p_{ij}^{(\text{state})} = \frac{e^{-\lambda_s d(i, j)}}{\sum_{r=1}^J e^{-\lambda_s d(i, r)}}. \quad (54)$$

**Definition D.5 (Query Distance).** Within a given query list  $l_i$  (with associated edge weights  $\{w_{i,1}, w_{i,2}, \dots, w_{i,k-1}\}$ ), the distance between queries at positions  $s$  and  $q$  is defined as

$$d_n(s, q) = \sum_{r=\min\{s, q\}}^{\max\{s, q\}-1} w_{i,r}, \quad (55)$$

with the convention that  $d_n(s, s) = 0$ .

**Definition D.6 (Query Transition Probability).** Let  $\lambda_n$  be a local sensitivity parameter. Then, for a given query list  $l_i$ , the probability of transitioning from the query at position  $s$  to the query at position  $q$  is given by

$$p_{sq}^{(\text{query})} = \frac{e^{-\lambda_n d_n(s, q)}}{\sum_{t=1}^k e^{-\lambda_n d(s, t)}}. \quad (56)$$

Before proceeding, we relabel the query lists as follows. Suppose that the first query from the  $(J+1)$ th list is observed in one of the initial  $J$  lists; then we designate that list as  $l_1$ . The remaining lists are then relabeled as  $l_2, l_3, \dots, l_J$  in order according to their proximity (as measured by  $d(i, 1)$ ) to  $l_1$ .

Next, we define a composite state as an ordered pair  $(i, q)$ , where  $i \in \{1, 2, \dots, J\}$  indicates the query list, and  $q \in \{1, 2, \dots, k\}$  indicates the position within that list. We assume that a new query behavior always starts from a fixed initial composite state:

$$\pi^{(0)}(i, q) = \delta_{i1} \delta_{q1}, \quad (57)$$

where  $\delta$  is the Kronecker delta.

Then, we define the one-step transition probability from a composite state  $(i, q)$  to another composite state  $(j, s)$  as the product of the list-level and query-level transition probabilities:

$$P_{(i,q) \rightarrow (j,s)} = p_{ij}^{(\text{state})} \cdot p_{sq}^{(\text{query})}. \quad (58)$$

We note that under this formulation the probability of reaching any given composite state after a sequence of transitions reflects the likelihood that the observed query behavior is generated by a normal user. In particular, by assigning higher transition probabilities

to paths corresponding to smaller distances, the model implicitly favors query sequences that are more "normal."

**COROLLARY D.7.** *With the initial composite state fixed as  $(i_0, q_0) = (1, 1)$ , consider the query behavior as a stochastic process. Then the probability of reaching a composite state  $(i_K, q_K)$  after  $K$  transitions is given by*

$$\pi^{(K)}(i_K, q_K) = \sum_{(i_1, q_1), \dots, (i_{K-1}, q_{K-1})} \prod_{n=0}^{K-1} P_{(i_n, q_n) \rightarrow (i_{n+1}, q_{n+1})}, \quad (59)$$

*where the sum is taken over all possible sequences of intermediate composite states.*

By combining the list-level transition process with the local (within-list) query transition process, our composite model assigns a well-defined probability to the event that a query sequence (starting from a fixed initial query, e.g., the first query of  $l_1$ ) evolves through a series of transitions to reach a specified composite state  $(i, q)$ . This probabilistic framework not only captures the behavior of normal users but also underpins the ATOM mechanism, thereby providing strong interpretability to our attack detection strategy.