Exploring Explainable AI in Cybersecurity Using the Kairos Dataset

-Tejas Mhadgut

This Project Addresses Enhancing Model Transparency for Improved Intrusion Detection and Attack Investigation

GROUP

RIT | Rochester Institute of Technology

ENVIRONMENT SETUP!

Embed Graphs	✅
Create Database	✅
Prepare	✅
Train	❌
Test	?
Evaluation	?
Anomalous Queue	?
Attack Investigation	?

Contributions!

- Hierarchical XAI for Kairos
- TGNN-Explainer Feature Importance

LITERATURE REVIEW

1. Graph clustering with graph neural networks.[A. Tsitsulin, J. Palowitch, B. Perozzi, and E. Müller]
2. Unveiling Molecular Moieties through Hierarchical Graph Explainability.[P. Sortino, S. Contino, U. Perricone, and R. Pirrone]
3. Believe the HiPe: Hierarchical perturbation for fast, robust, and model-agnostic saliency mapping.[J. Cooper, O. Arandjelovic, and D. J. Harrison]
4. Everybody needs a little help: Explaining graphs via hierarchical concepts.[Under Review]
5. Explaining Temporal Graph Models through an Explorer-Navigator Framework.[W. Xia, M. Lai, C. Shan, Y. Zhang, X. Dai, X. Li, and D. Li]
6. Explainability Methods for Graph Convolutional Neural Networks.[P. E. Pope, S. Kolouri, M. Rostami, C. E. Martin, and H. Hoffmann]

PAPER REVIEW: Graph clustering with graph neural networks

OBJECTIVE: Introduction of DMoN (Deep Modularity Networks), a novel GNN-based graph clustering approach.

01

Addresses limitations of existing GNN pooling methods for graph clustering.

02

Directly optimizes modularity, a metric measuring the strength of a graph's division into clusters.

03

Resolves cluster collapse, where all nodes are assigned to one cluster.

04

Incorporates a null model to prevent collapse and ensure meaningful clustering.

05

Relevant for Further Research into Hierarchical XAI.

RIT | Rochester Institute of Technology

PAPER REVIEW: Believe the HiPe – Hierarchical Perturbation for Saliency Mapping

OBJECTIVE: Introduces HiPe, a model-agnostic technique for saliency mapping in AI models.

01

Primarily used to explain predictions in image classification tasks using Perturbations.

02

Starts with large overlapping regions, observes their effect on model predictions.

03

Refines by subdividing into smaller regions for detailed analysis.

04

Captures salient features across scales, from coarse to fine-grained details.

05

Offers insights for applying hierarchical methods in understanding model decisions in cybersecurity contexts.

RIT | Rochester Institute of Technology

PAPER REVIEW: Everybody Needs a Little HELP – Explaining Graphs via Hierarchical Concepts

OBJECTIVE: **HELP method provides explanations for how graph-based AI models (GNNs) make decisions.**

01

Analyzes patterns in node behavior after applying GNN layers.

02

Groups similar nodes into clusters.

03

Merges connected nodes in a cluster into “super-nodes”, calculating their properties by averaging.

04

Reveals how simpler components form complex structures step by step..

05

Enhances understanding of model decision-making at multiple levels.

RIT | Rochester Institute of Technology

Challenges for Hierarchical XAI in Cybersec

- Limited availability of hierarchical XAI Research Papers, especially for GNNs.
- Existing work predominantly focuses on image datasets.
- Many relevant papers are still under review, restricting established resources.

Approach 1: Adapting HiPe for Hierarchical Perturbation on Kairos

PROPOSED METHOD: Apply hierarchical perturbation to the provenance graph in Kairos.

01

Edge Removal: Simulate absence of specific events by selectively removing edges to evaluate their importance in anomaly detection or model performance.

02

Node Removal: Simulate absence of entities by removing nodes, highlighting their overall contribution.

03

Timestamp Modification: Assess model sensitivity to event timing by altering timestamps.

04

Hierarchical Integration: Group these perturbations into a hierarchical framework to create an explainable hierarchy.

Approach 2: Adapting HELP for Hierarchical Explanations on Kairos

PROPOSED METHOD: **Based on the HELP paper's hierarchical graph explanation framework.**

01

Segment the Provenance Graph: Divide into 15-minute time windows for temporal analysis.

02

Apply GNN Layers: Transform node features into embeddings capturing temporal and structural relationships within each window.

03

Cluster Nodes: Use methods like DMoN to group nodes based on behavioral and structural similarities.

04

Form Concept Nodes: Merge connected components in each cluster into higher-level concept nodes.

05

Analyze Concept Nodes: Calculate reconstruction error of edges associated with concept nodes across time windows. Use this to evaluate concept stability and significance over time.

TGNN-Explainer

OBJECTIVE: **Explain predictions made by Temporal Graph Neural Networks (TGNNs) in dynamic graphs, focusing on the temporal dependencies of events.**

01

Explorer: Uses Monte Carlo Tree Search (MCTS) to identify event combinations contributing to predictions.

02

Navigator:A pre-trained feed-forward network that guides the Explorer by predicting important events, reducing the search space.

03

Navigator Inputs: Node Representations, Edge features, and Temporal Features.

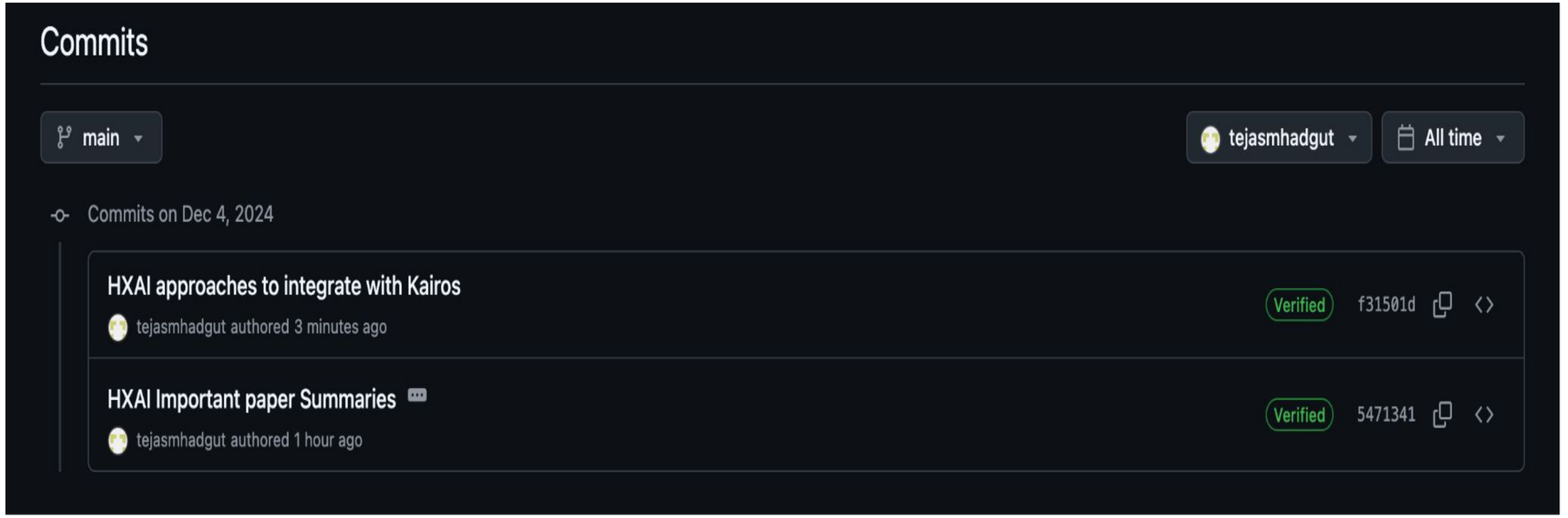
04

Provides Instance level Explanations.

05

Evaluation Metrics is Fidelity and Sparsity.

Commits


Commits
main
tejasmhadgut
All time
Commits on Dec 4, 2024
HXAI approaches to integrate with Kairos
tejasmhadgut authored 3 minutes ago
Verified
f31501d
HXAI Important paper Summaries
tejasmhadgut authored 1 hour ago
Verified
5471341

THANK YOU TEAM!