Explainable Malware Detection via Relational Graph Neural Networks with Bidirectional Relations

Monday Onoja, Zekeri Adams, Peter Anthony, Martin Homola

Department of Applied Informatics, Faculty of Mathematics Physics and Informatics Comenius University in Bratislava, Slovakia

COMENIUS UNIVERSITY BRATISLAVA

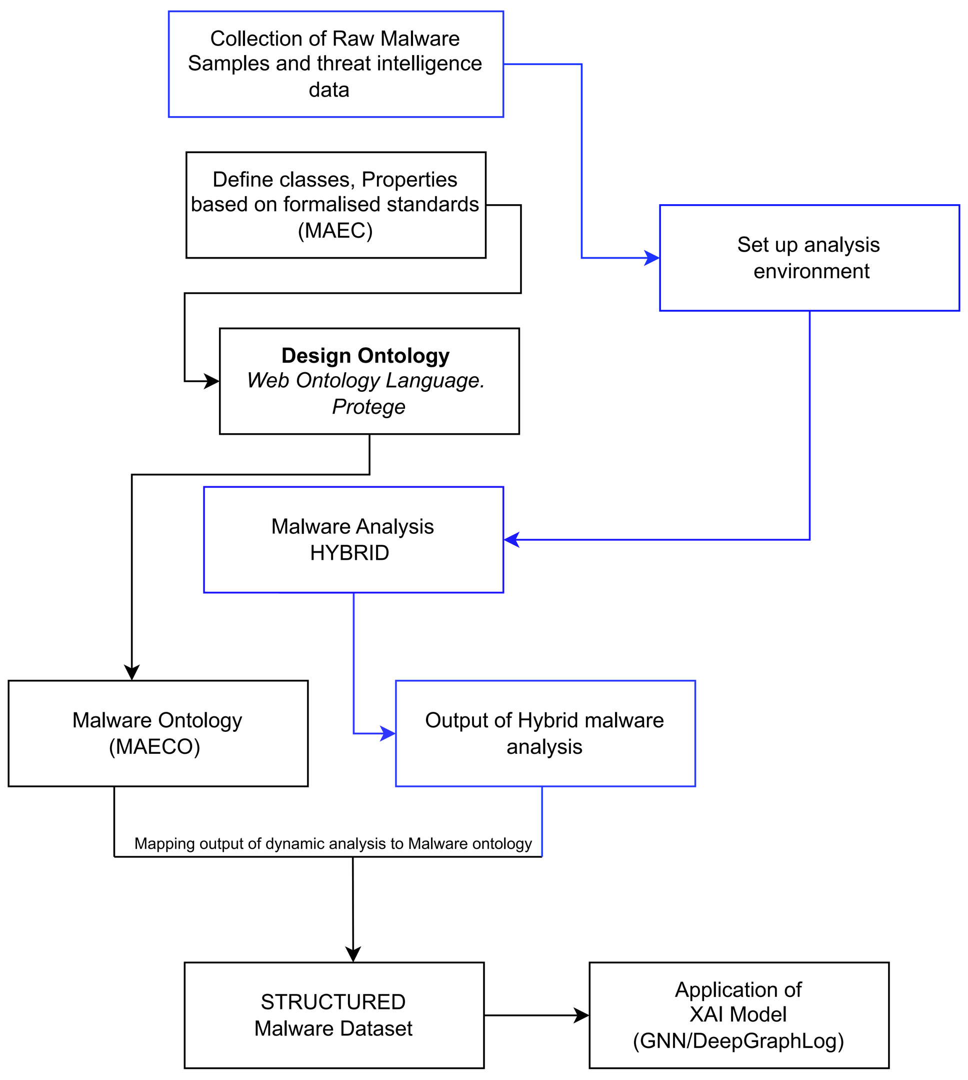
Motivation

- Ontology provides a formal framework for representing malware concepts and relationships in both machine- and human-readable forms, enabling improved detection and explainability (Švec et al., 2024).
- Integrating dynamic malware attributes into ontology enhances discrimination and interpretability by capturing a wider range of behaviors and expressing them through clear, standardized vocabularies (Amita Dessai, 2021; Owoh et al., 2024).
- Combining ontology-driven relational data with Graph Neural Networks (GNNs) produces malware detection systems that are not only accurate but also transparent and robust, offering deeper semantic understanding and improved explainability (Bilot et al., 2023; Shokouhinejad et al., 2025)

Problem statement/Gap

- Existing ontology-based malware models (e.g., Anthony et al., 2023; Mojžiš et al., 2023) make progress in interpretability but fail to exploit the relational structure inherent in malware data.
- The PEMalware Ontology (Švec et al., 2024) focuses solely on static features, which are vulnerable to obfuscation and evasion, limiting detection robustness.
- Graph Neural Networks (GNNs) offer strong performance for malware detection but remain black-box models, lacking the transparency needed for analyst trust and forensic insight.

Project Work Flow



Figur 1: Project workflow
Figure 1 presents the entire Project workflow, including Definiton of classes, Ontology design, Malware Analysis, Ontology-based datset construction and application of XAI model.

MAECO Core Classes and Relationships

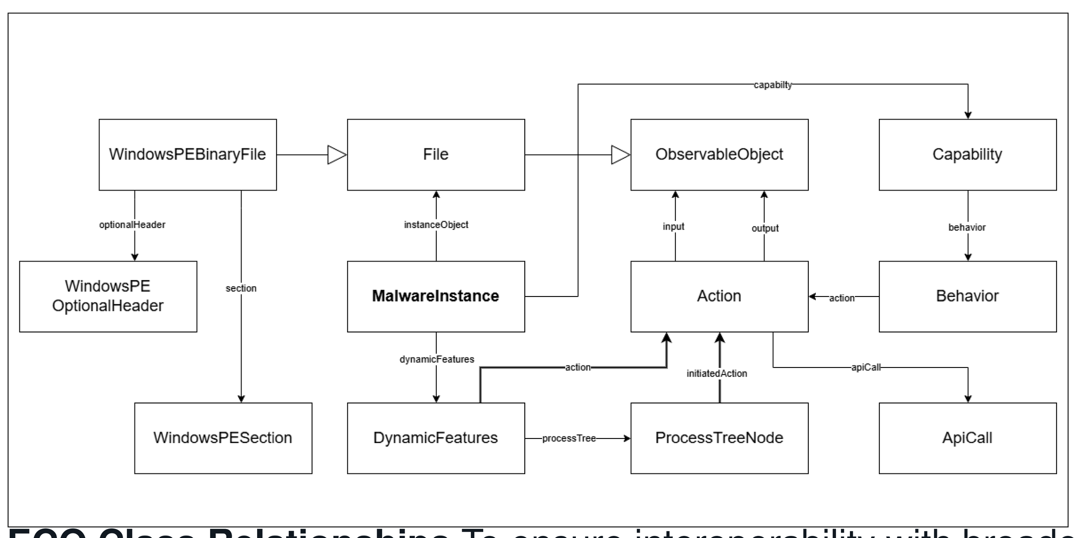


Figure 2: MAECO Class Relationships To ensure interoperability with broader cyber threat intelligence standards, MAECO establishes explicit semantic links between MAEC objects and STIX Cyber-Observable Objects (SCOs).

Proof of Concept: GCN and RGCN

To demonstrate the suitability of GNN on Ontology-based dataset, we constructed a Pytoch Geometric data (PyG) suitable for GNN from the knowledge graphs constructed by Trizna et al. (2024) derived from the existing PE Malware ontology constructed by ˇSvec et al. (2022) from the EMBER dataset (based on static features) with 1000 label(2) samples. In the first phase (Table 1) of the experiments, we used only the numeric feature subset of the dataset to test the effectiveness of edge reversal. while in the second phase (Table 2) we used the whole feature set to train the RGCN model and node-level and graph-level explaination with captum explainer (Table 3)

Result 1

Table1: Performance comparison of GCN and RGCN models with and without edge reversal
GCN1: GCN without edge reversal, GCN2: GCN with edge reversal, RCGN1: RGCN without edge reversal, RGCN2: RGCN with edge reversal.

Model	Precision	Recall	F1-score	Accuracy	TPR
GCN1	64	79	71	67	55
GCN2	78	46	58	67	87
RGCN1	72	55	65	67	74
RGCN2	99	97	98	98	98

Result 2: RGCN2 with Captum explainer

Table2: Performance of RGCN2 on full feature set

Model	Precision	Recall	F1-score	Accuracy	TPR
RGCN2	82	85	84	82	85

Table 3: Fidelity and Relative Drop Metrics

Metric	Value
Mean fidelity$^-$	0.1016
Mean fidelity$^+$	0.8698
Mean relative drop	0.1302

Node - levle and graph Level Explaination

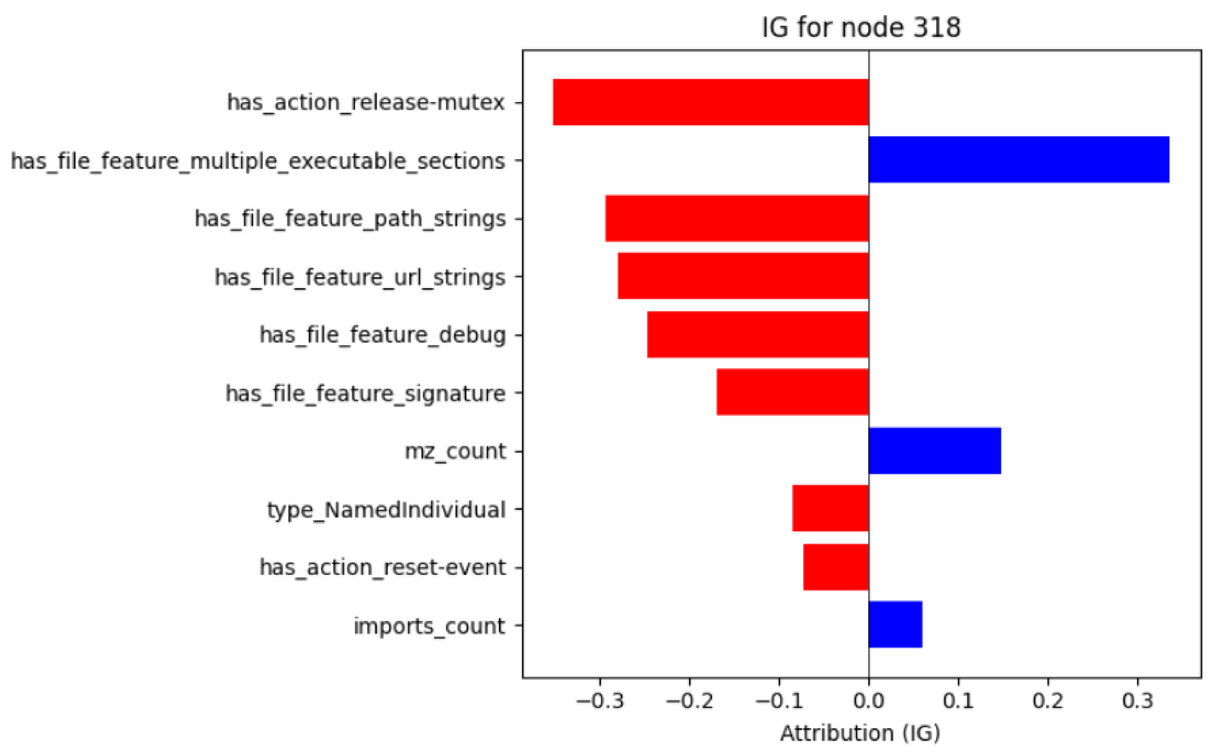


Figure 3: Node Level (Local) Explaination for Node 318

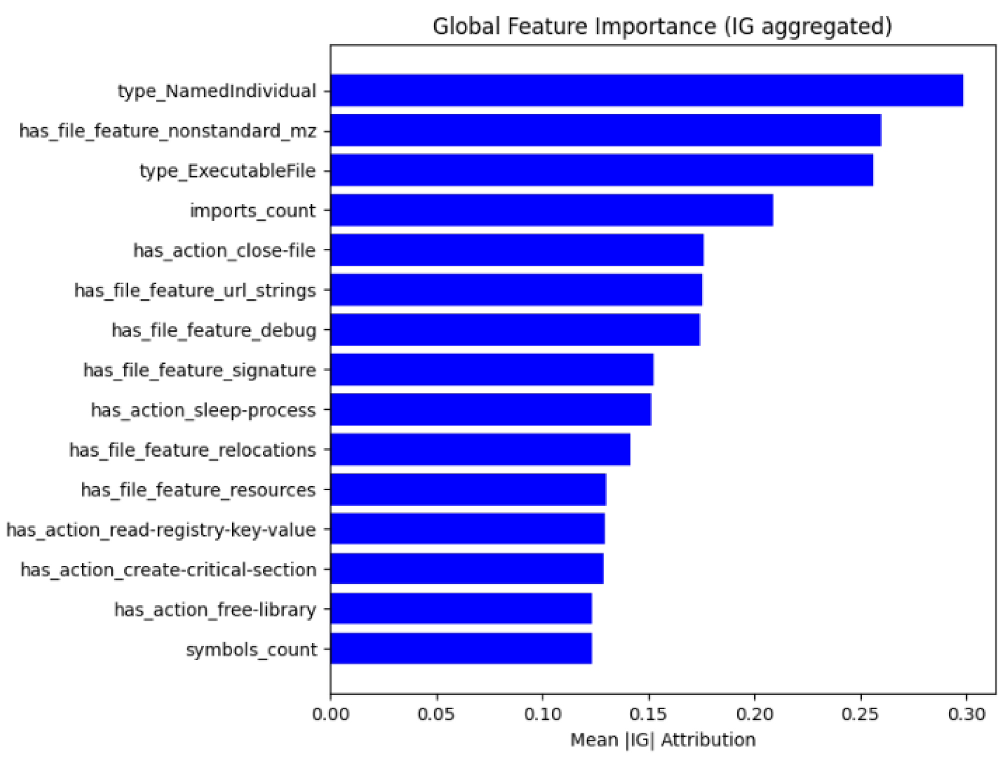


Figure 4: Graph Level Explaination (Global)

Discussion and Conclusion

The experimental results demonstrate that the Relational Graph Convolutional Network (RGCN), particularly when enhanced with edge reversal, effectively captures complex relational and semantic dependencies within ontology-based malware data. Using Captum's Integrated Gradients (IG), both node- and graph-level explanations were generated to interpret the model's predictions. The node-level attributions revealed that distinctive file-level features such as `has_file_feature_multiple_executable_sections`, `has_action_release-mutex`, and `has_file_feature_path_strings` strongly indicate malicious behavior, whereas attributes like `imports_count` and `mz_count` occasionally acted as counter-signals.
- At the global level, the explanations confirm that RGCN2 relies primarily on ontology-derived semantic relations rather than raw numeric attributes, reinforcing the advantage of integrating relational reasoning into malware detection models. Overall, these findings demonstrate that ontology-based relational learning not only enhances classification accuracy but also enables transparent, human-interpretable insights, establishing a solid foundation for the neuro-symbolic extensions to be explored in the future.

References

Scan for reference list

ACKNOWLEDGMENT: This work is funded by the EU NextGenerationEU through the Recovery and Resilience Plan for Slovakia under project No. 09105-03-V02-00064.

Scan to send a mail
Email: `monday.onoja@fmph.uniba.sk`

Presented at the 22nd International Conference on Principles of Knowledge Representation and Reasoning - Melbourne

MATFYZ CONNECTIONS