#95 Explaining Provenance-Based GNN Detectors with Interpretable Graph Structural Features

Dr. Kunal Mukherjee

Kunal Mukherjee*, Joshua Wiedemeier[$], Tianhao Wang[$], Muhyun Kim[$], Feng Chen[$], **Murat Kantarcioglu***, and Kangkook Jee[$]

***Virginia Tech**
[$]University of Texas at Dallas

Dr. Murat Kantarcioglu

Agenda

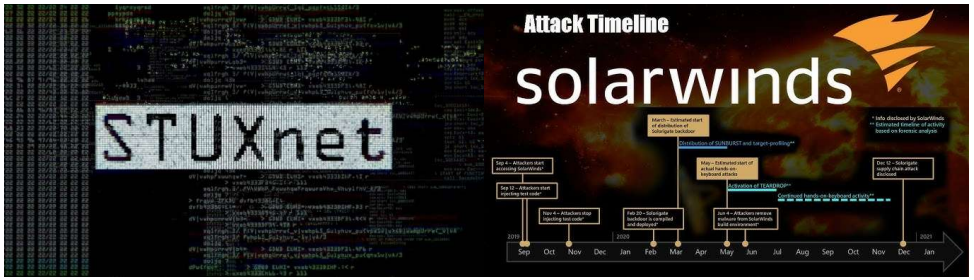
1. Background
2. Motivation
3. **GNN-based IDS Explainability**: ProvExplainer
4. Future Work
5. Conclusion

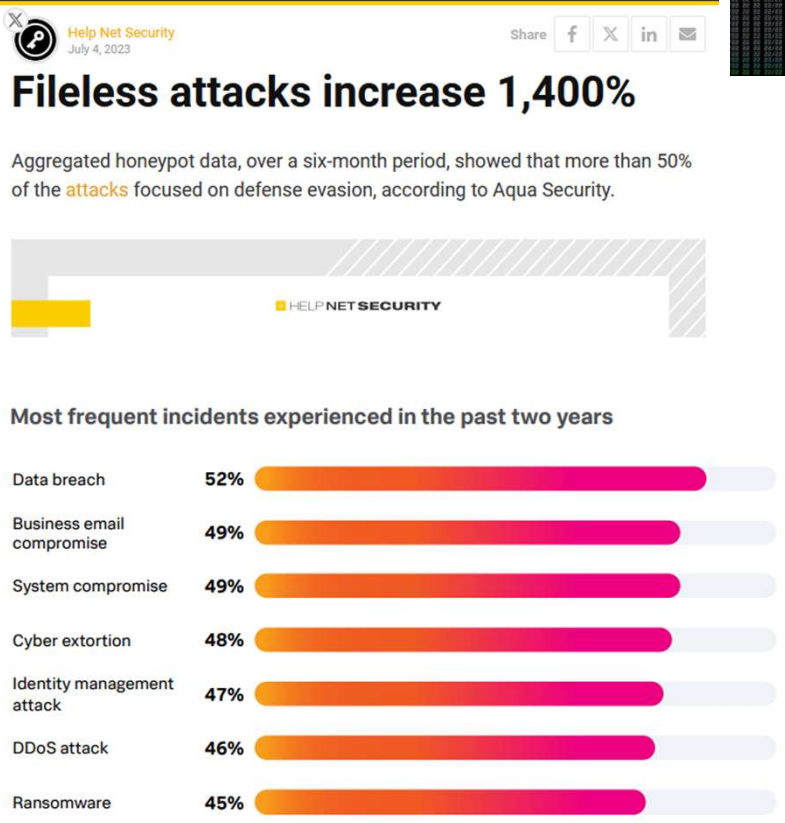
Agenda

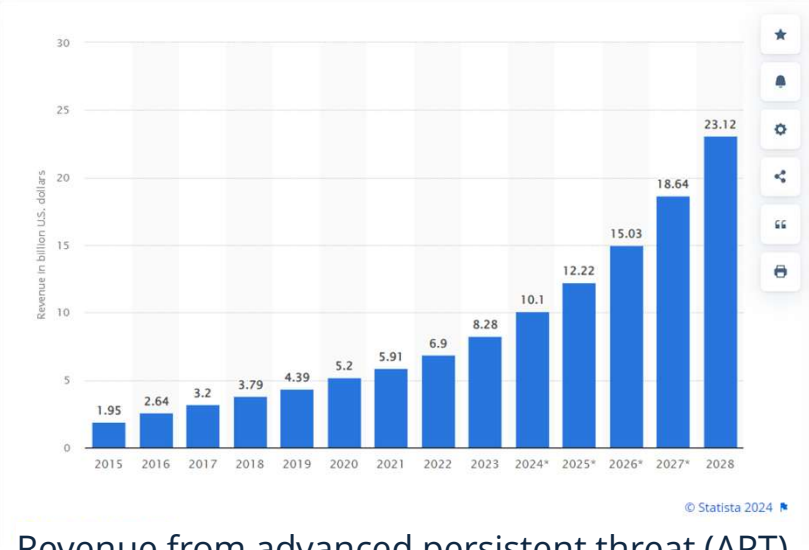
1. **Background**
2. Motivation
3. GNN-based IDS Explainability: ProvExplainer
4. Future Work
5. Conclusion

Background:

Stealthy Attacks





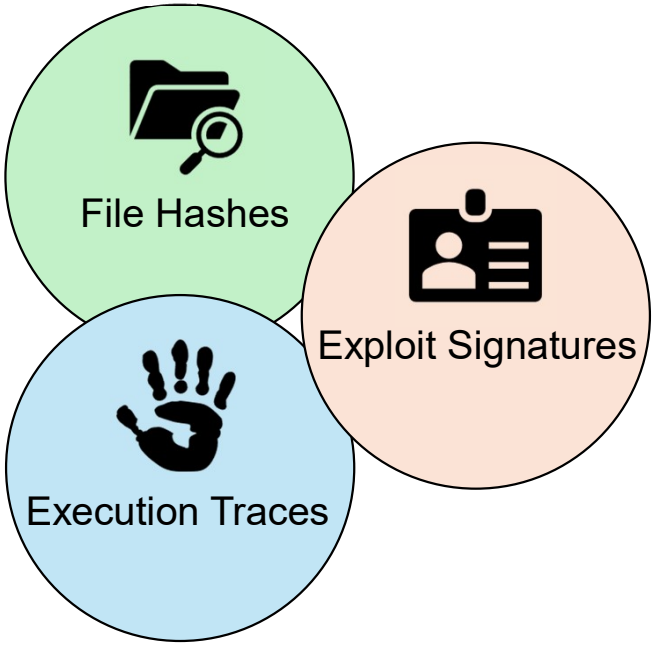


Revenue from advanced persistent threat (APT) from 2015 to 2027

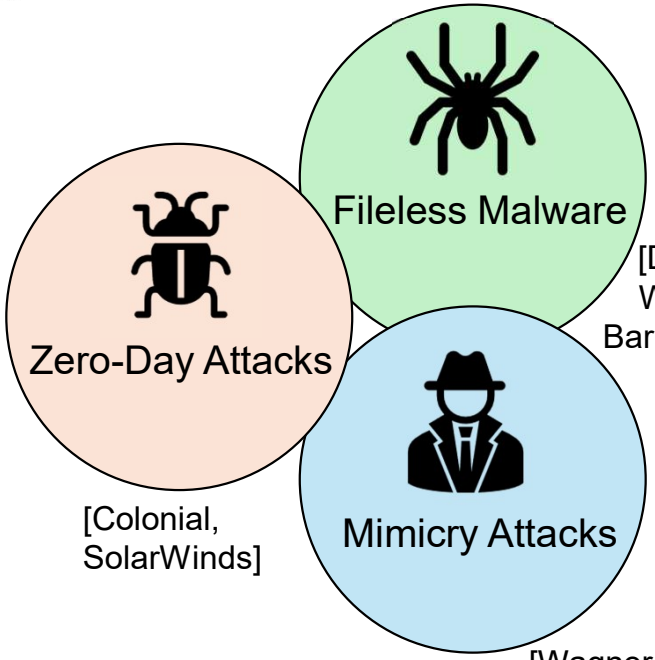
Background:

Static Host Defenses

Traditional Host *Intrusion Detection System (IDS)* detects **static artifacts**



Adversaries evade detection with **stealthy techniques**



[Dang '19, Wang '20, Barr-Smith '21]

[Colonial, SolarWinds]

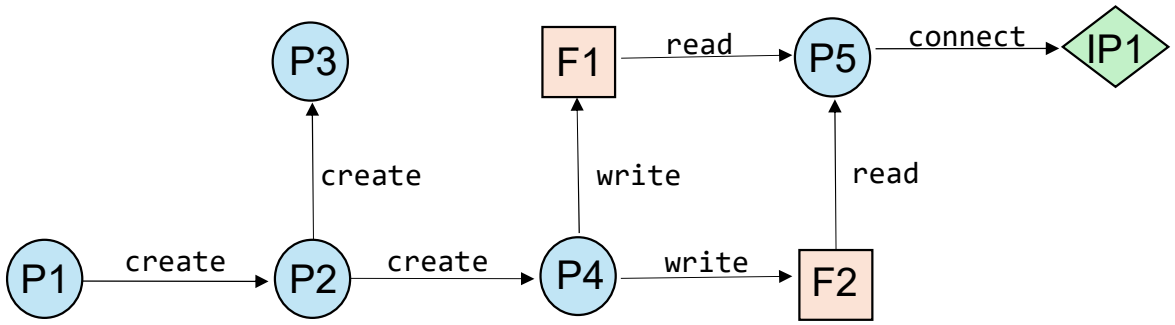
[Wagner & Soto '02, Tan & Maxion '03]

Background:

System Provenance

- **System Provenance** championed as a *host-based* dynamic defense
  - Influential works [Hassan '19, Wang '20, Han ‘21, Rehman ’24, Goyal ‘24]

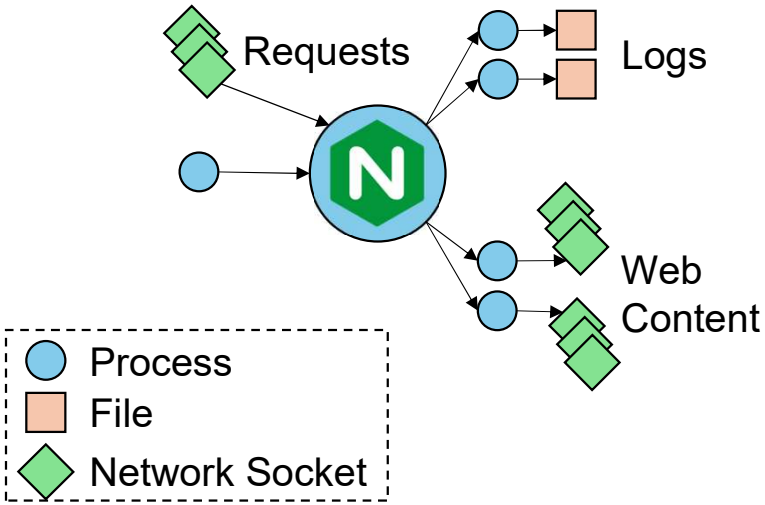
- System Provenance *causally* connects system resources
  - Captures *dynamic* control and data dependencies



How can System Provenance help detect stealthy attacks?

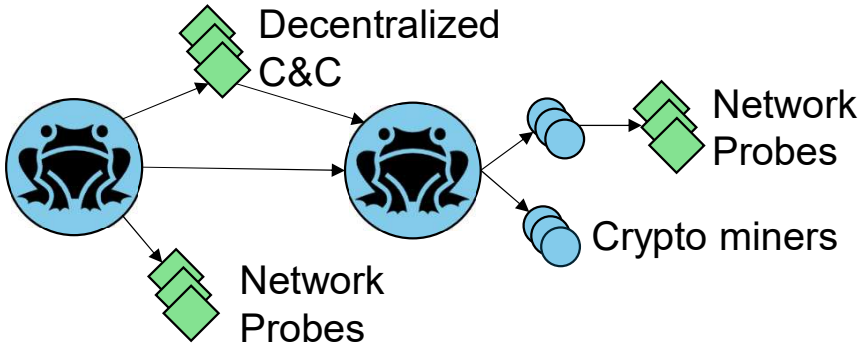
Background:

Fileless Malware Example



The FritzFrog botnet has claimed **over 20,000 victims**

Exploits Log4j and weak SSH passwords to impersonate vulnerable Nginx webservers



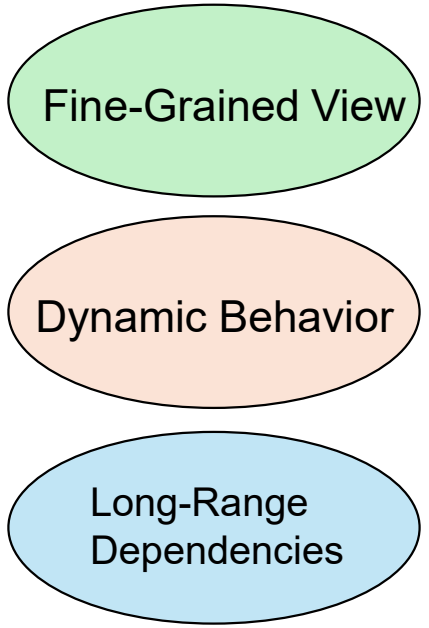
Background:

GNN-based Intrusion Detection System (IDS)

GNN-based IDS Prediction Granularity

Velickovic '17 ⇨ Graph-level detection

Zeng ‘22, Cheng ‘24, Rehman ’24 ⇨ Interaction-level detection



Primary Roadblock to Provenance-Based IDS Adoption

Trust in Provenance-based IDS (PIDS) has **not been established**

Explanation of alerts and detection have **not been verified**

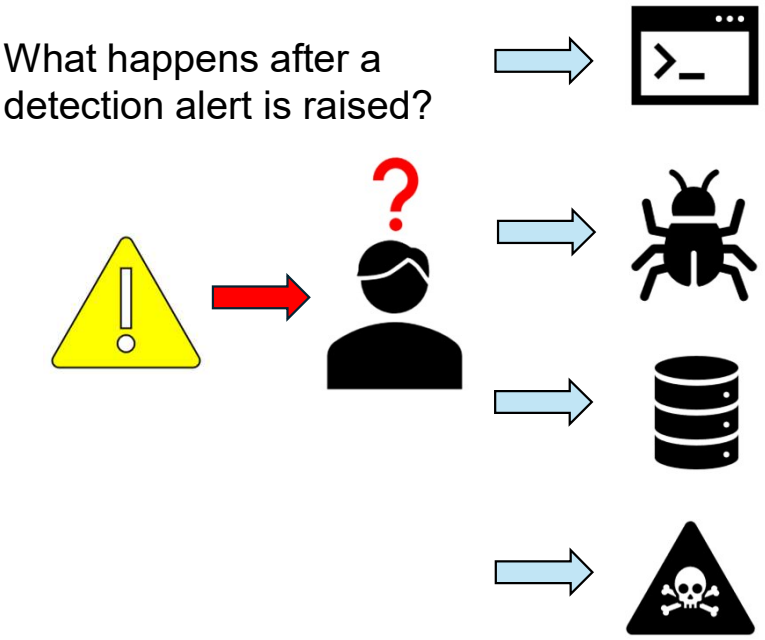
Generic Explanation Harness is required for GNN-based PIDS

Agenda

1. Background
2. **Motivation**
3. GNN-based IDS Explainability: ProvExplainer
4. Future Work
5. Conclusion

Motivation:

Explainability of Provenance-Based IDS



Trust in Provenance-based IDS has **not been established**

Human operators need **justification** and **guidance** to triage and respond to alerts

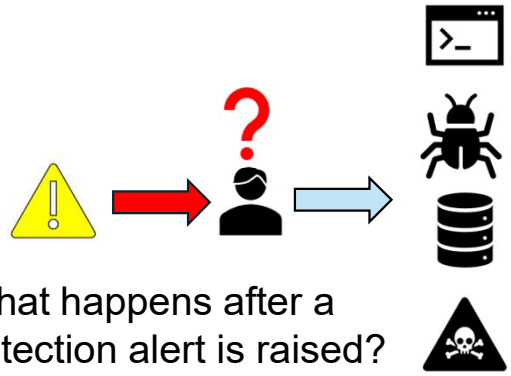
Agenda

1. Background
2. Motivation
3. **GNN-based IDS Explainability: ProvExplainer**
4. Future Work
5. Conclusion

Kunal Mukherjee, and et. al. "Explaining Provenance-Based GNN Detectors with Interpretable Graph Structural Features ," PST, 2026.

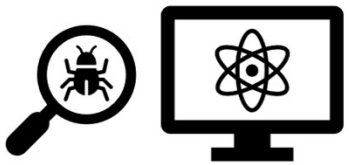
GNN-based IDS Explainability: ProvExplainer

Motivation

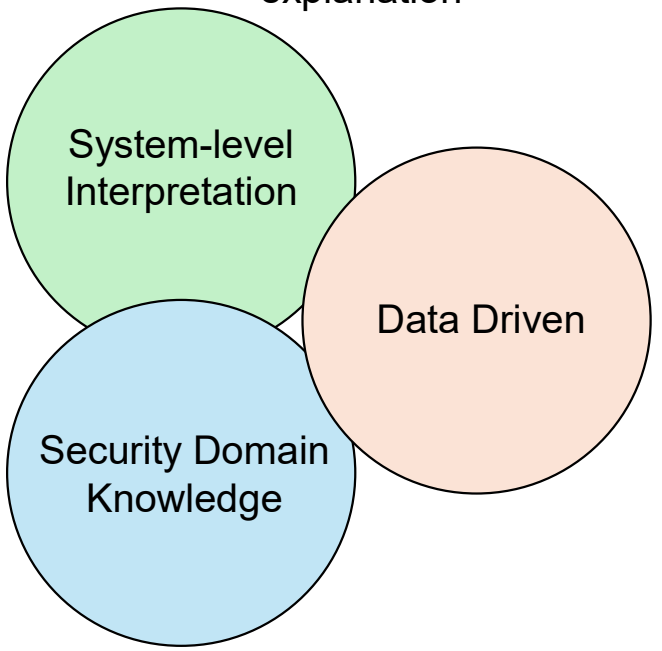


What happens after a detection alert is raised?

Security-aware explanations to **interpret** and bring **context** to alerts.



Desired qualities of security-aware explanation



Motivation : *GNN-based IDS*

SOTA provenance–based IDS [Cheng ‘24, Rehman ’24] use GNN graph embeddings for detection

Explanation verification is hindered due to **black-box** nature of GNN

General purpose GNN Explainers are **not** security-aware

Design

ProvExplainer:
Security Aware
Explanation Framework

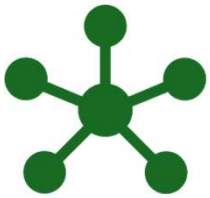
Extract Security Aware Features

Train Surrogate DT

Interpret Surrogate DT

GNN-based IDS Explainability: ProvExplainer

Design



Instance-level

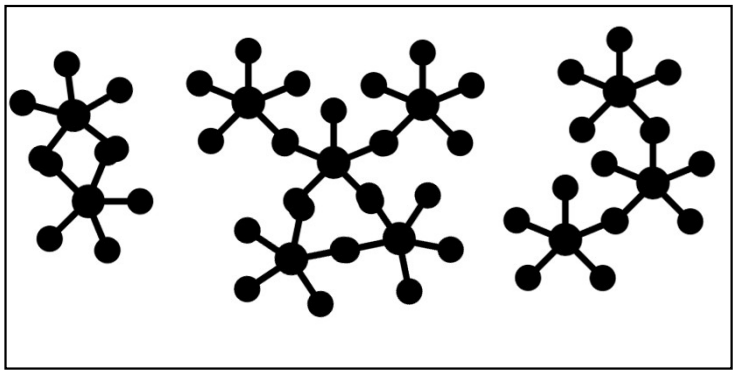
Black box

Structure only

GNN-based IDS Explainability: ProvExplainer

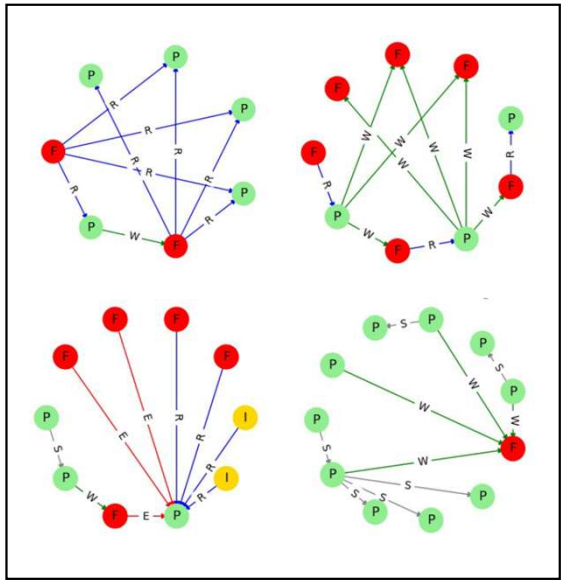
Design : *Security-Aware Graph Structural Features*

Anomalous Graphs



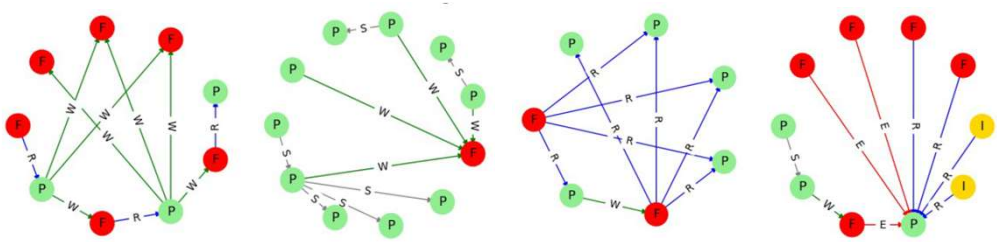
Subgraph Pattern Mining [Graph Evolutionary Rule Miner]

Subgraph Patterns



GNN-based IDS Explainability: ProvExplainer

Design : *Security-Aware Graph Structural Features*



Pattern **Size** – total number of nodes

Pattern **Diversity** – average "importance score" of node types

Pattern **Support** – average number of occurrences across *all* graphs in the dataset

Pattern **Coverage** – total number of occurrences across *anomalous* graphs in the dataset

Discriminant Pattern Score = α * **size** + β * **diversity** + γ * **support** + η * **coverage**

GNN-based IDS Explainability: ProvExplainer

Design : *Security-Aware Graph Structural Features*

Discriminant Pattern Score = α * **size** + β * **diversity** + γ * **support** + η * **coverage**

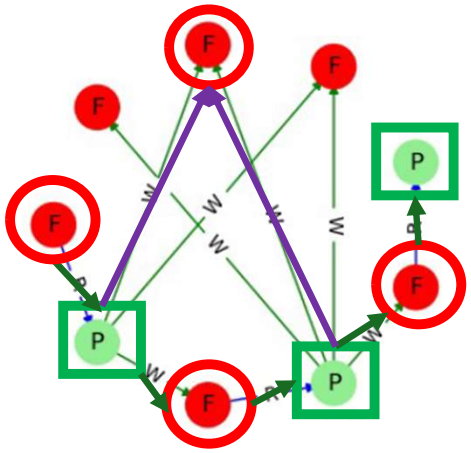
Hyperparameter: α=0.1, β=0.2, γ=0.4,and η= 0.3

Prioritize **support** and **coverage** as the primary differentiators between patterns followed by **diversity** and pattern **size**

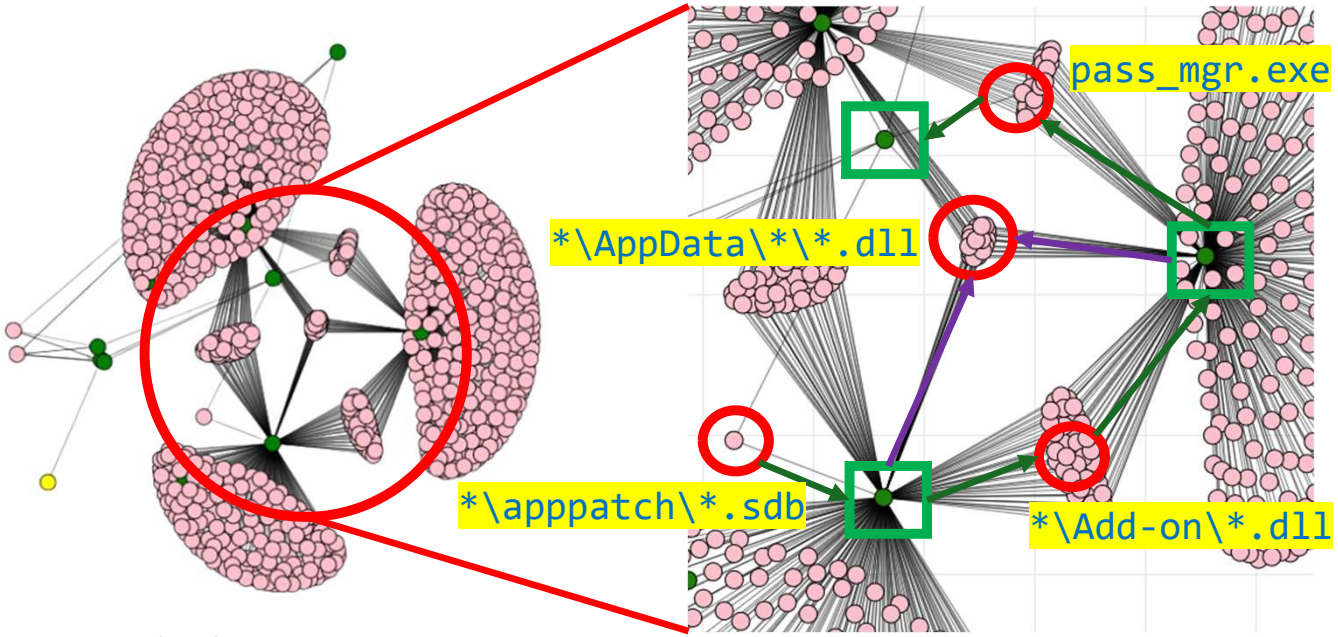
Select the patterns that have the **highest** Discriminant Pattern Score

GNN-based IDS Explainability: ProvExplainer

Design : *Example of Explanations*



Example of patterns selected for explanation



Exploits *firefox.exe* by downloading a malicious password manager, enabling it to read and store sensitive system files.

GNN-based IDS Explainability: ProvExplainer

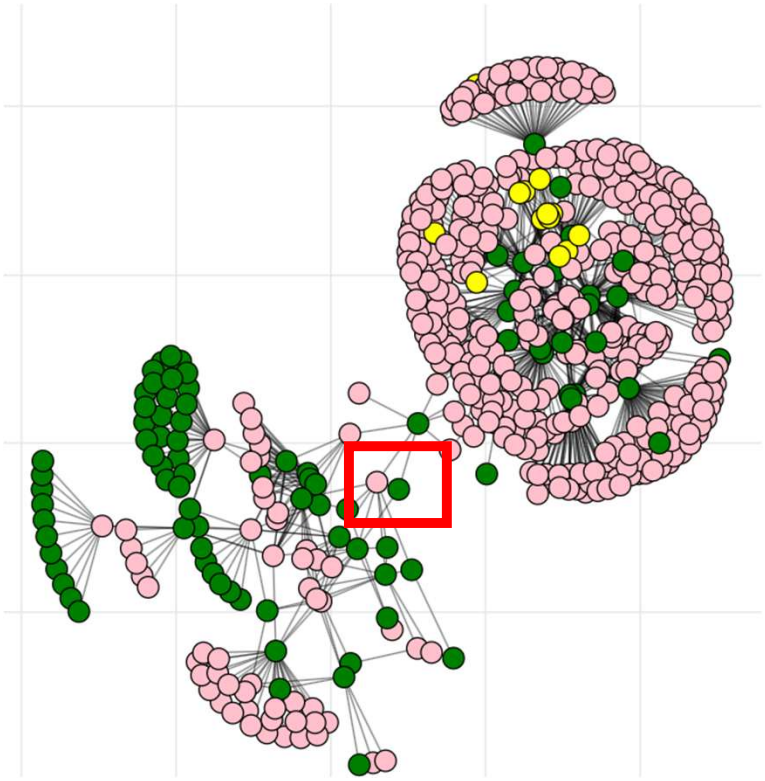
Design : *Graph Structural Features*

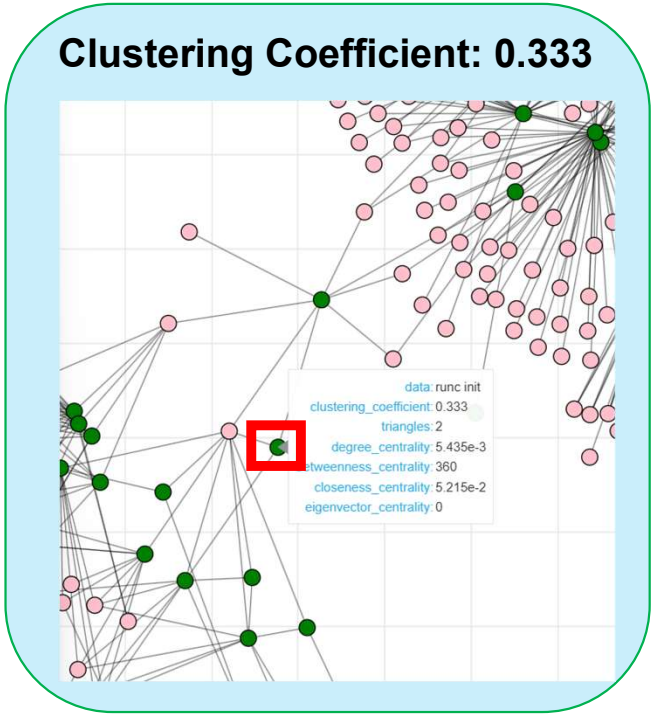
Clustering Measures
Clustering Coefficient
Triangles

Centrality Measures	
Degree	Closeness
Betweenness	Eigenvector

GNN-based IDS Explainability: ProvExplainer

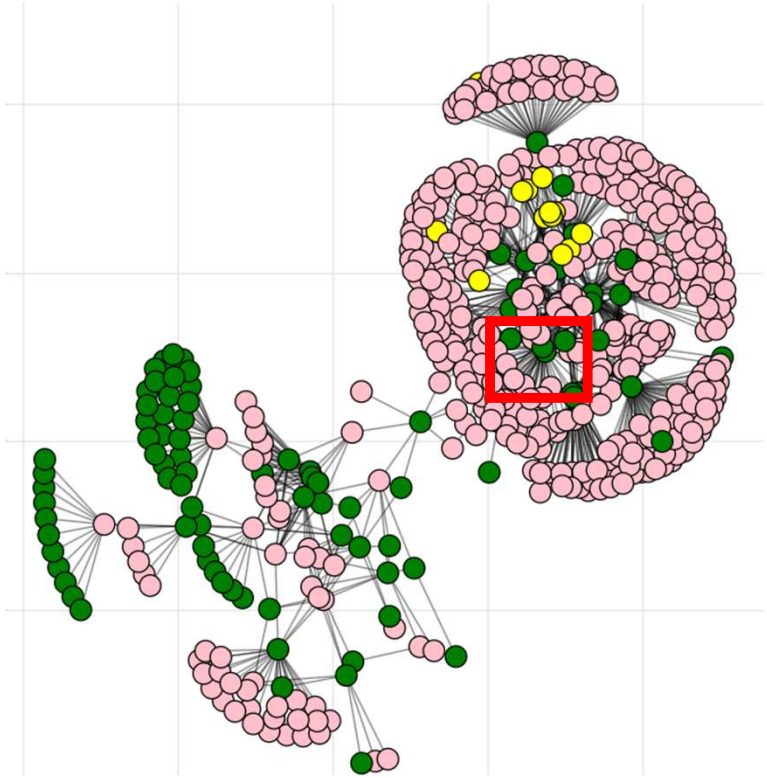
Design : *Graph Structural Features*



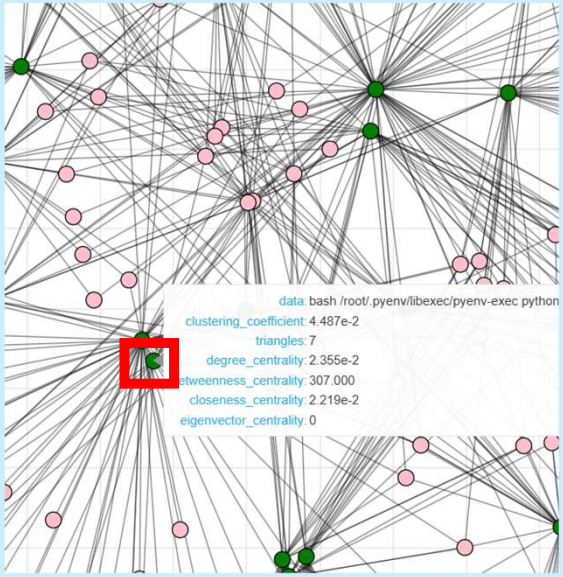


GNN-based IDS Explainability: ProvExplainer

Design : *Graph Structural Features*

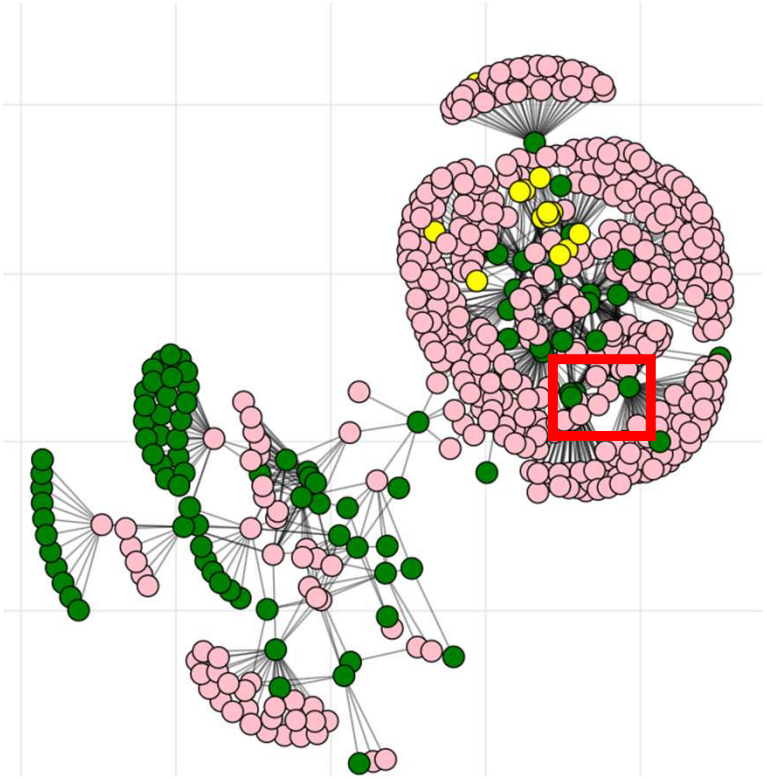


Clustering Triangles: 7

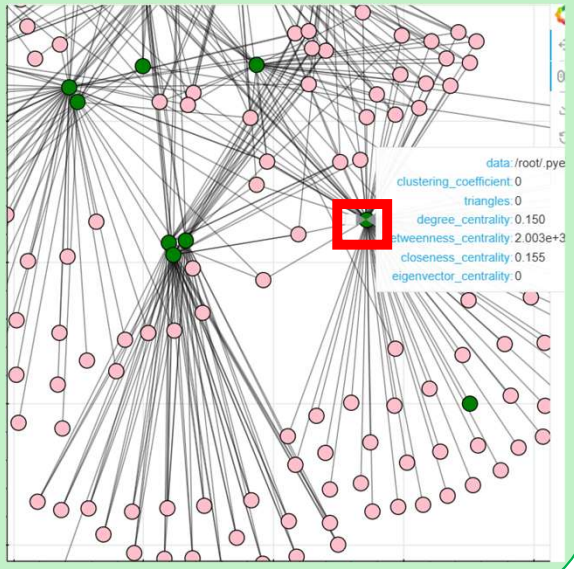


GNN-based IDS Explainability: ProvExplainer

Design : *Graph Structural Features*

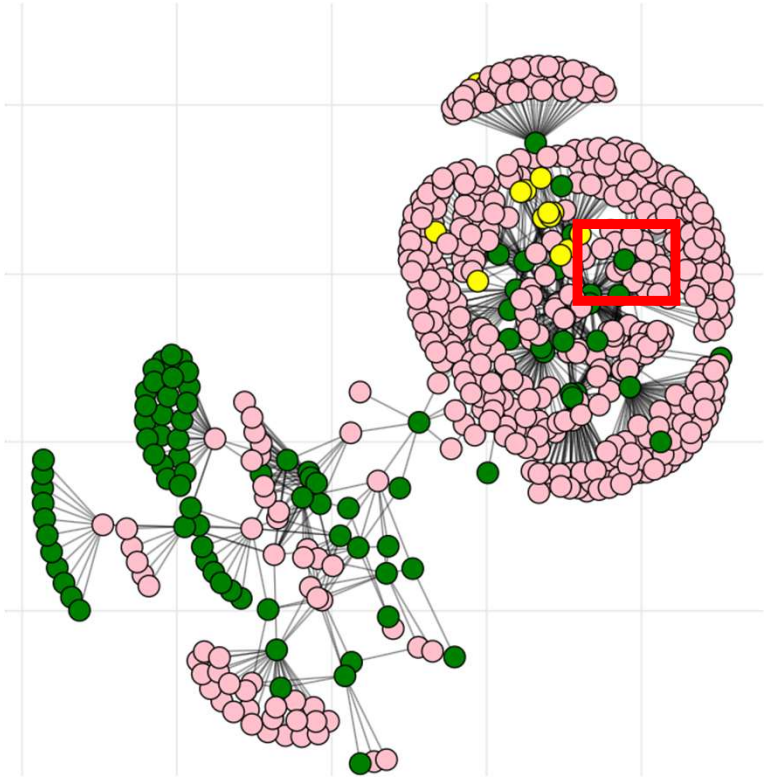


Degree Centrality: 0.150

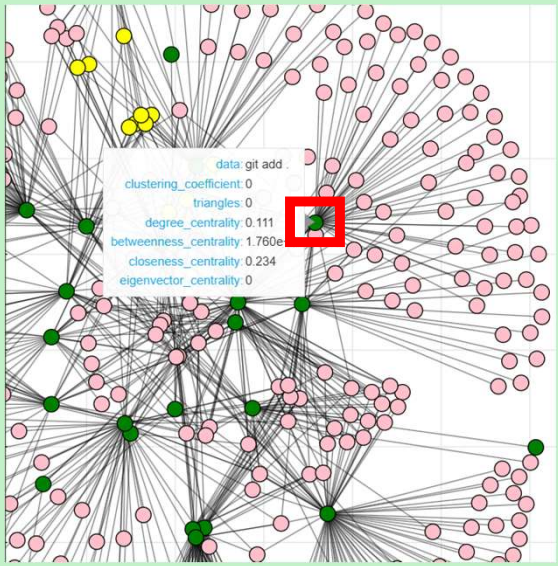


GNN-based IDS Explainability: ProvExplainer

Design : *Graph Structural Features*

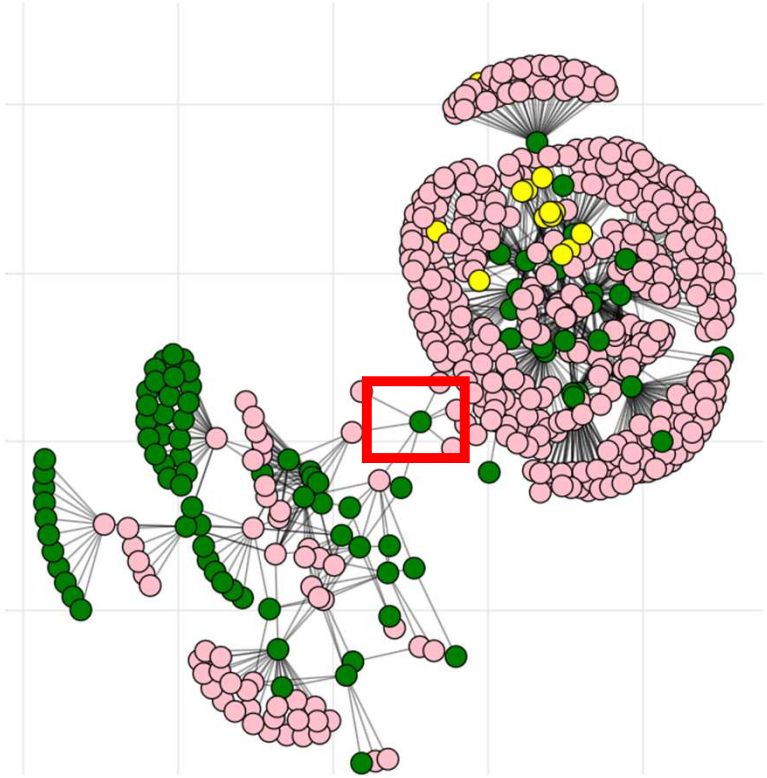


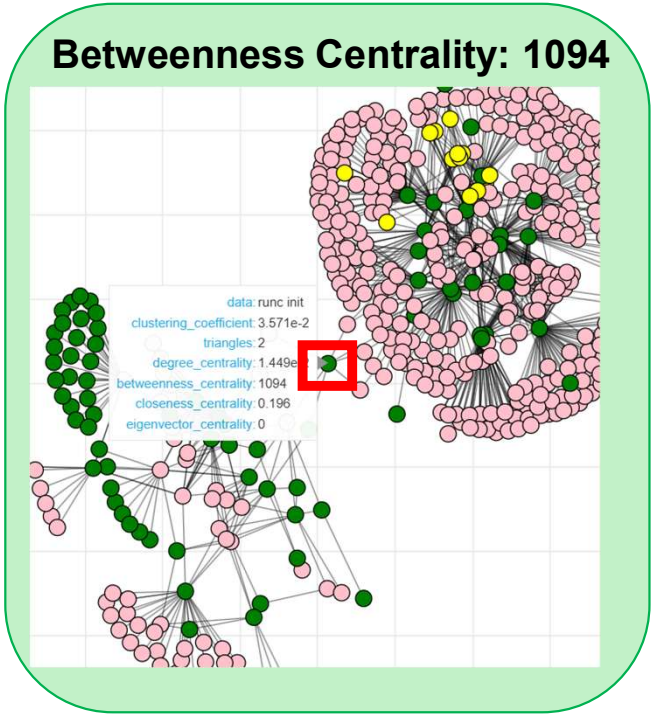
Closeness Centrality: 0.234



GNN-based IDS Explainability: ProvExplainer

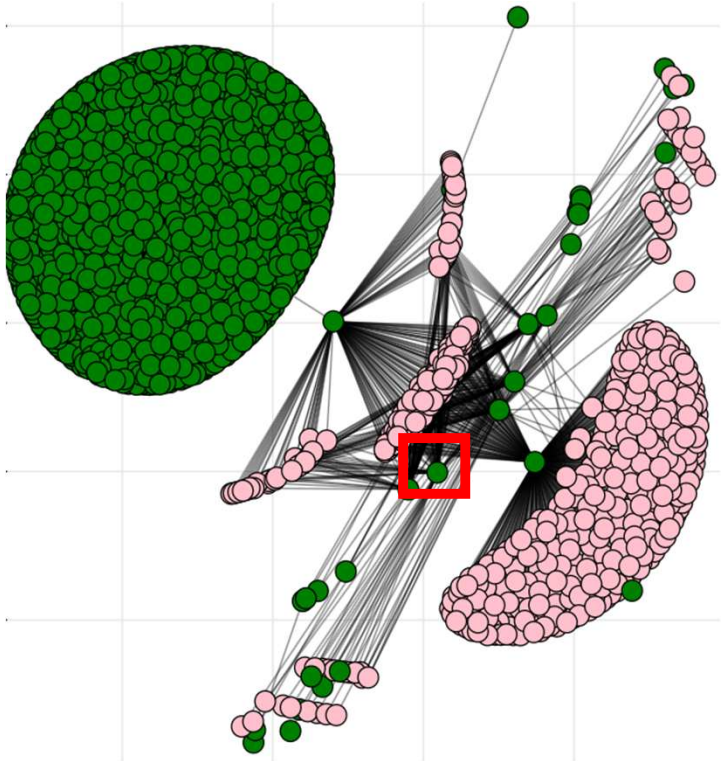
Design : *Graph Structural Features*



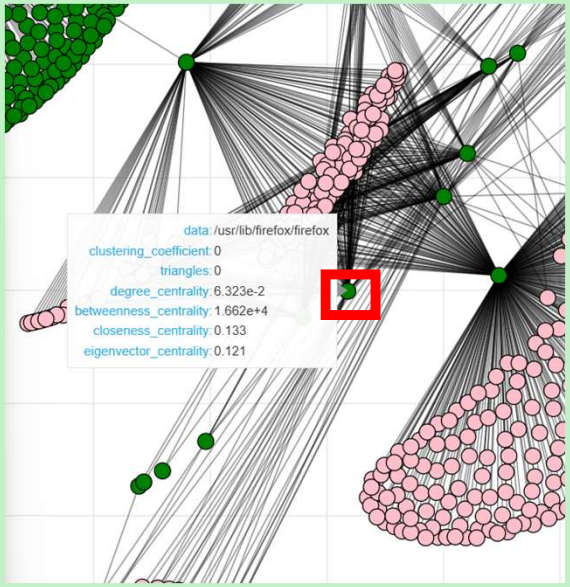


GNN-based IDS Explainability: ProvExplainer

Design : *Graph Structural Features*



Eigenvector Centrality: 0.121



GNN-based IDS Explainability: ProvExplainer

Design : *Using the Graph Structural Features*

Train a surrogate DT using the count of discriminant subgraph patterns and avg. of graph structural features

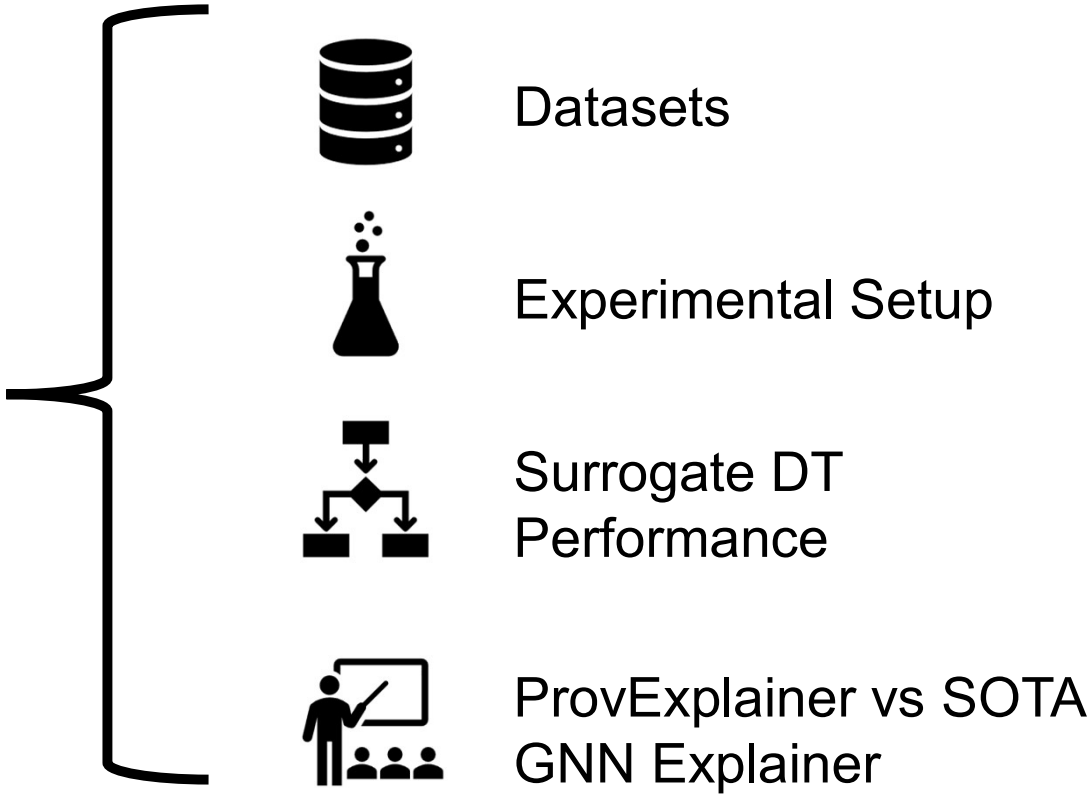
Use the surrogate DT to classify graphs

Extract node information from patterns along the decision path

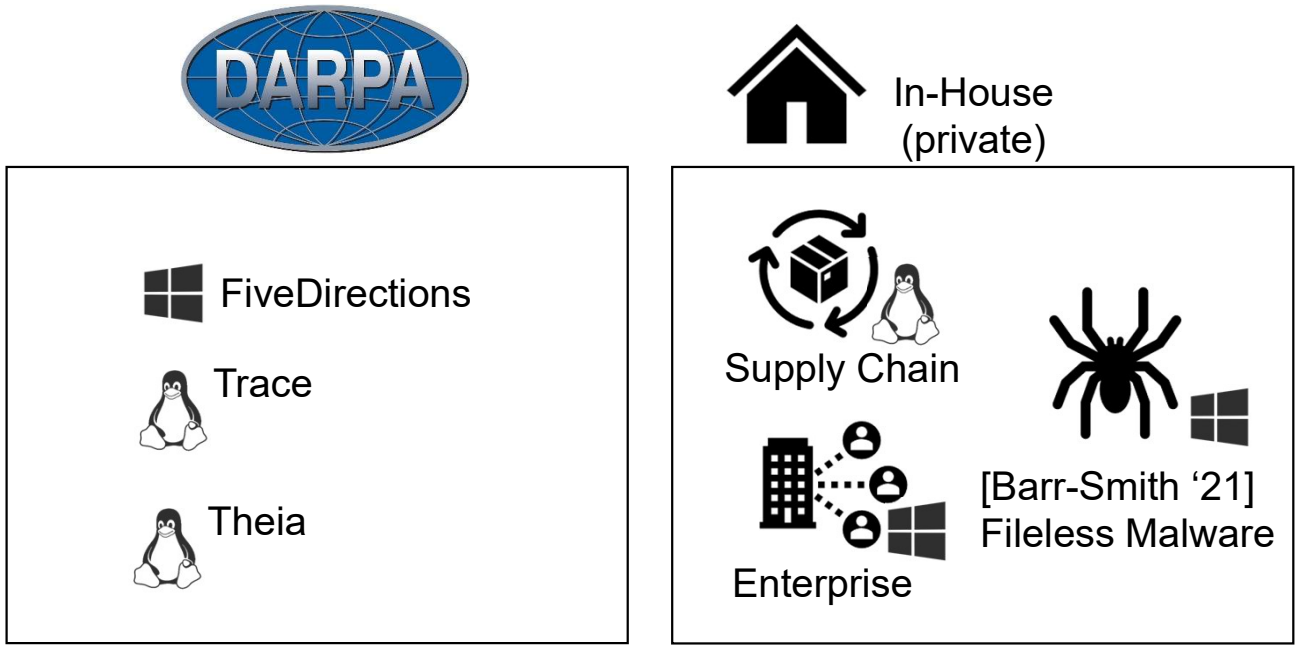
Patterns that appear **more often** in the decision path are deemed **more** important

Evaluation

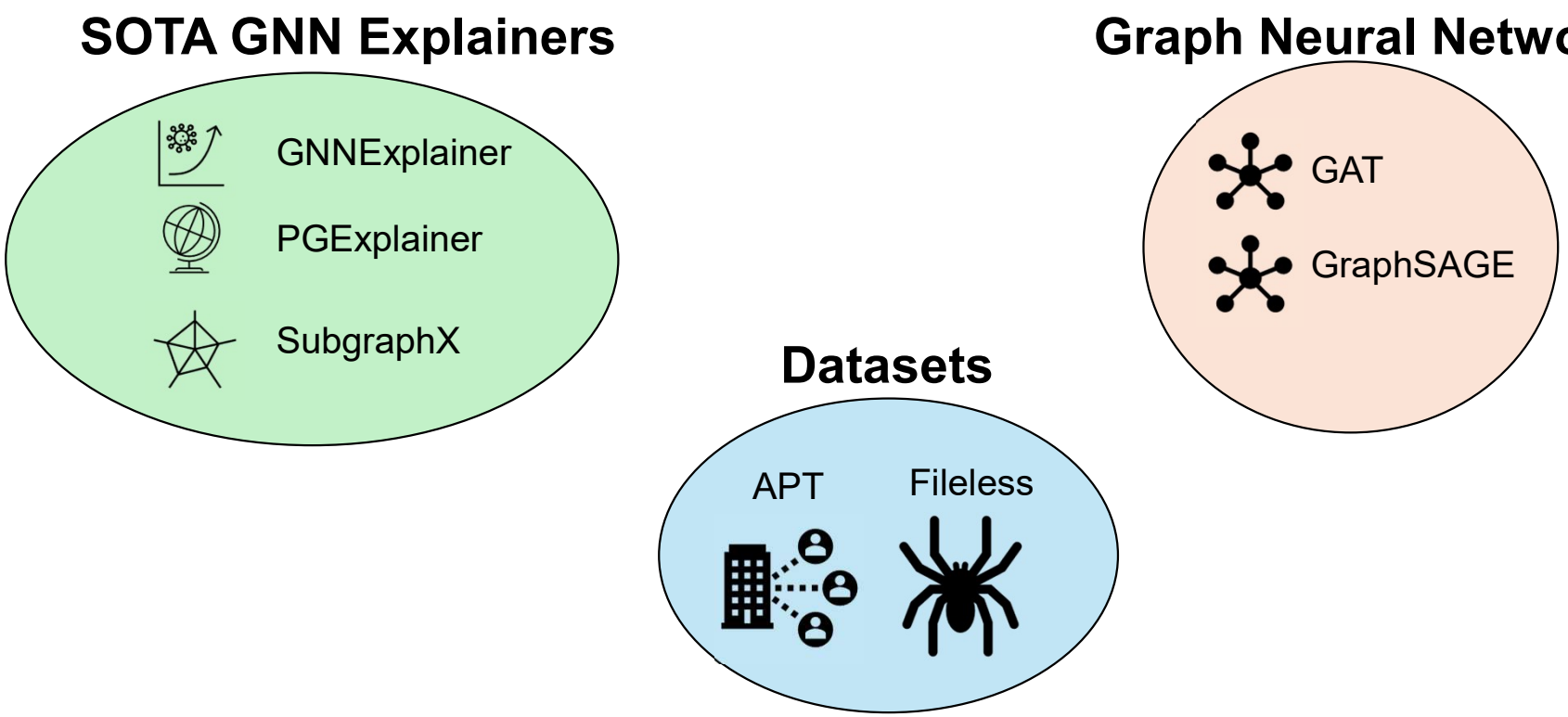
Evaluation



Evaluation : *Anomaly Detection Datasets*

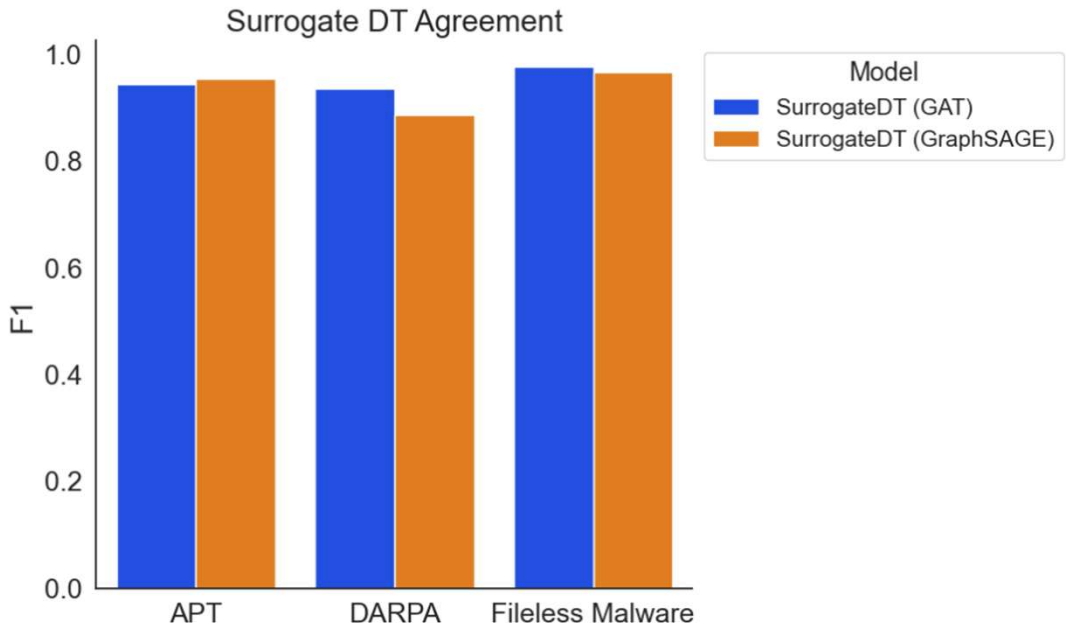


Evaluation : *Experimental Setup*



GNN-based IDS Explainability: ProvExplainer

Evaluation : *Surrogate DT Performance*



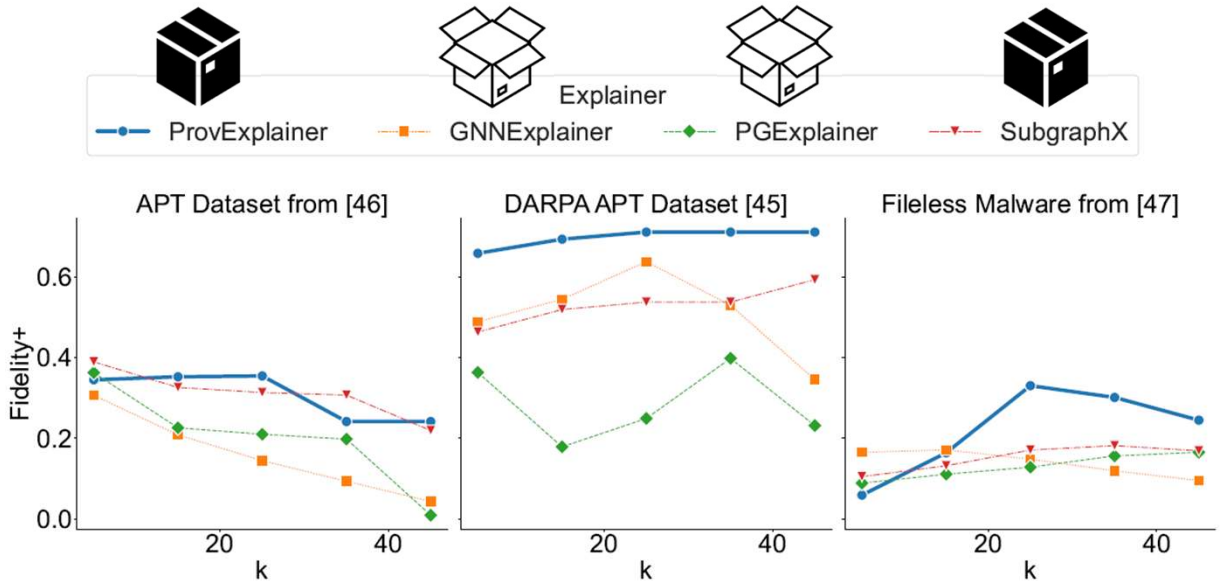
Surrogate DTs have high agreement with GNNs

GNN-based IDS Explainability: ProvExplainer

Evaluation : *SOTA Comparison – Fidelity+*

Fidelity+ = the difference in prediction after removing **k important** nodes from the graph

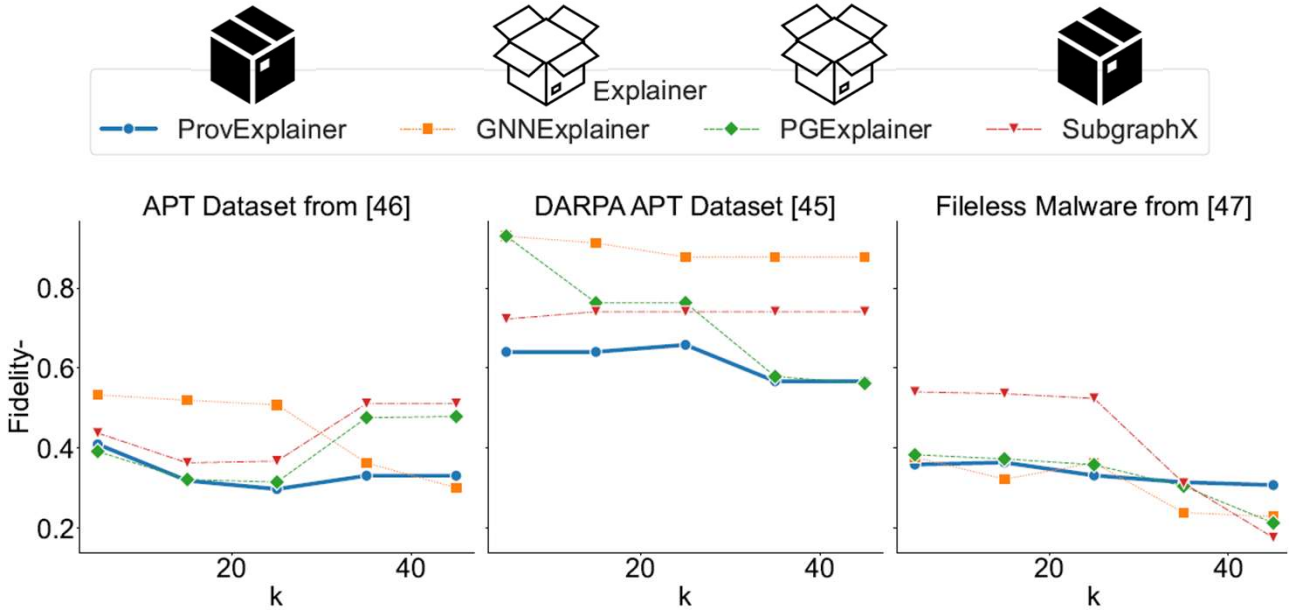
(**higher** is better)



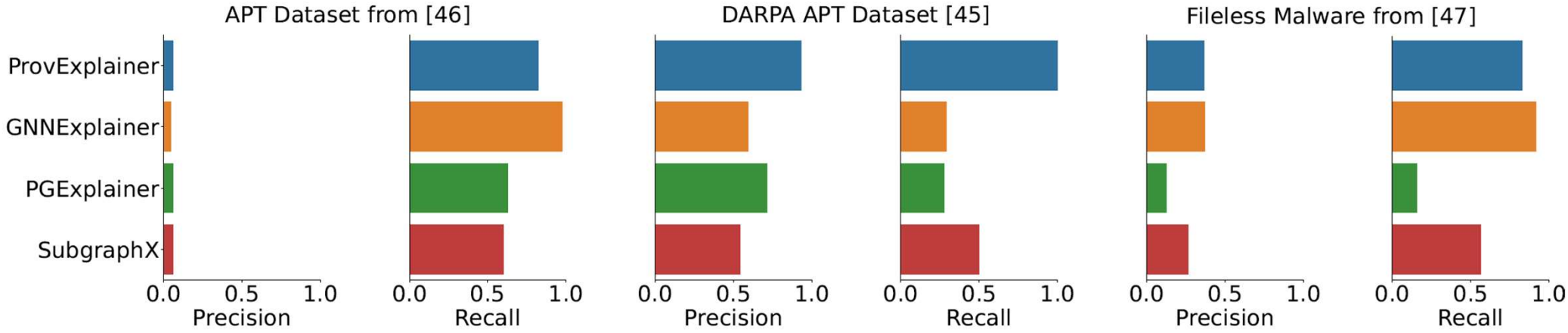
Evaluation : *SOTA Comparison – Fidelity-*

Fidelity- = the difference in prediction after removing **k unimportant** nodes from the graph

(**lower** is better)



Evaluation : *SOTA Comparison – Precision/Recall*



Competitive performance compared to SOTA

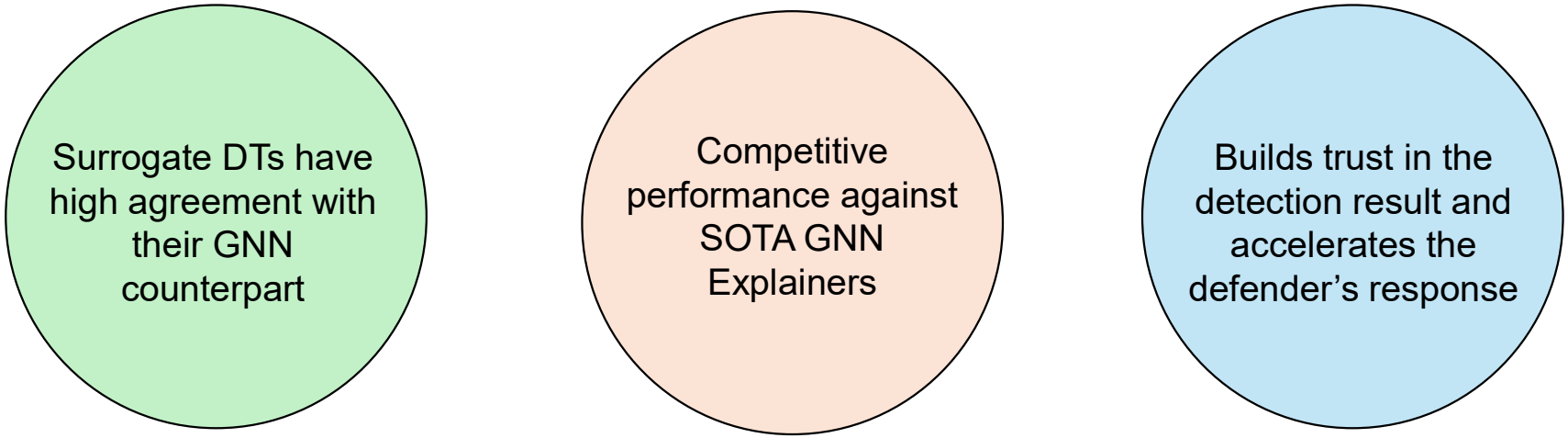
Extracts important system entities

Superior performance in DARPA dataset

GNN-based IDS Explainability: ProvExplainer

Project Summary

ProvExplainer generates **security-aware explanations**



Agenda

1. Background
2. Motivation
3. GNN-based IDS Explainability: ProvExplainer
4. **Future Work**
5. Conclusion

Future Works

1. **Robust and generalizable explanations across domains**
   Study how well explanation quality holds under noisy data, adversarial manipulation, and distribution shift, while extending the methods to other high-stakes graph applications beyond intrusion detection.

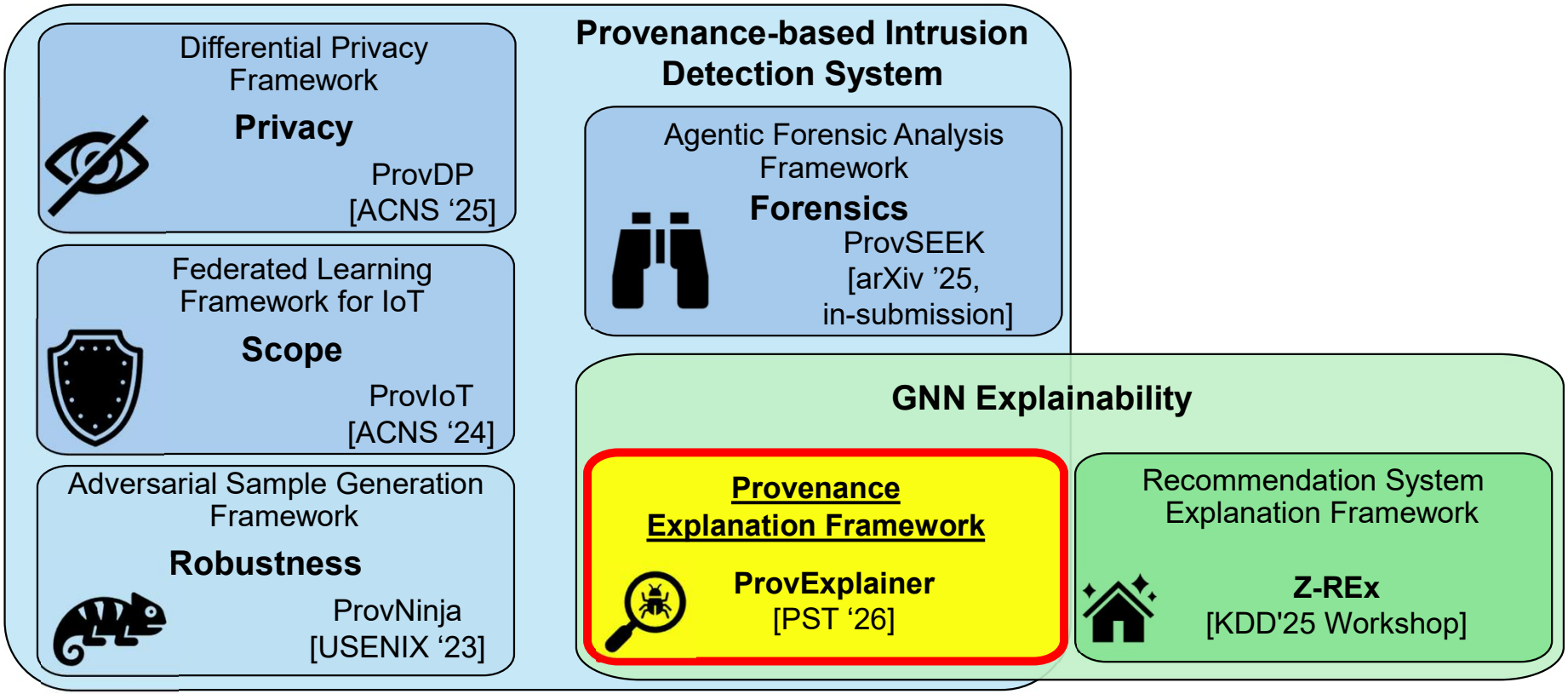
2. **Temporal and dynamic graph explainability**
   Incorporate time-aware explanations so the frameworks can explain not only *which* nodes and edges mattered, but also *when* and *in what sequence* interactions contributed to the decision.

3. **Counterfactual and actionable explanations**
   Extend the current frameworks beyond identifying important structures to generating *actionable counterfactuals*, such as what minimal graph changes would alter an IDS alert.

Agenda

1. Background
2. Motivation
3. GNN-based IDS Explainability: ProvExplainer
4. GNN-based RecSys Explainability: Z-REx
5. Future Work
6. **Conclusion**

Human-Interpretable ML Explanations in High-Stakes Graph ML



THANK YOU

Looking forward to your questions and comments!

Scan the QR code to access the paper

ProvExplainer
[PST ‘26]
[Extended Version]

KUNAL MUKHERJEE

Website

LinkedIn

www.KUNMUKH.com

kunmukh@gmail.com

Research Papers

[1] **Kunal Mukherjee**, Joshua Wiedemeier, Tianhao Wang, James Wei, Feng Chen, Muhyun Kim, Murat Kantarcioglu, and Kangkook Jee. "Evading Provenance-Based ML Detectors with Adversarial System Actions," in *32nd USENIX Security Symposium (USENIX Security 23)*, 2023, pp. 1199–1216.2.

[2] **Kunal Mukherjee**, Zachary Harrison, and Saeid Balaneshin. "Z-REx: Human-Interpretable GNN Explanations for Real EstateRecommendations," (Oral) KDD Workshop ML on Graphs in the Era of Generative AI (MLoG-GenAI), 2025.

[3] **Kunal Mukherjee**, Joshua Wiedemeier, Qi Wang, Junpei Kamimura, John Junghwan Rhee, James Wei, Zhichun Li, Xiao Yu, Lu-An Tang, Jiaping Gui, and Kangkook Jee. "ProvIoT: Detecting Stealthy Attacks in IoT through Federated Edge-Cloud Security," in *International Conference on Applied Cryptography and Network Security*, Springer, 2024, pp. 241–268.

[4] **Kunal Mukherjee**, Jonathan Yu, Partha De, and Dinil Mon Divakaran. "ProvDP: Differential Privacy for System Provenance Dataset," in *International Conference on Applied Cryptography and Network Security*, Springer, 2025, pp. 189-219.

[5] **Kunal Mukherjee**, Joshua Wiedemeier, Tianhao Wang, Muhyun Kim, Feng Chen, Murat Kantarcioglu, and Kangkook Jee."Explaining Provenance-Based GNN Detectors with Interpretable Graph Structural Features," in *22nd Annual International Conference on Privacy, Security, and Trust* (PST 2025).

[6] **Kunal Mukherjee**, Murat Kantarcioglu. “LLM-driven Provenance Forensics for Threat Investigation and Detection,” arXiv preprint arXiv:2508.21323, 2025.

References

Wagner & Soto ’02 - Wagner, David, and Paolo Soto. "Mimicry attacks on host-based intrusion detection systems." *Proceedings of the 9th ACM Conference on Computer and Communications Security*. 2002.

Tan & Maxion ’03 - Tan, Kymie MC, and Roy A. Maxion. "Determining the operational limits of an anomaly-based intrusion detector." *IEEE Journal on selected areas in communications* 21.1 (2003): 96-110.

Velickovic ’17 - Veličković, Petar, et al. "Graph attention networks." *arXiv preprint arXiv:1710.10903* (2017).

Hassan ’19 - Hassan, Wajih Ul, et al. "Nodoze: Combatting threat alert fatigue with automated provenance triage." *network and distributed systems security symposium*. 2019.

Dang ’19 - Dang, Fan, et al. "Understanding fileless attacks on linux-based iot devices with honeycloud." *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services*. 2019.

Ying ’19 - Ying, Zhitao, et al. "Gnnexplainer: Generating explanations for graph neural networks." *Advances in neural information processing systems* 32 (2019).

Wang ’20 - Wang, Qi, et al. "You Are What You Do: Hunting Stealthy Malware via Data Provenance Analysis." *NDSS*. 2020.

Han ‘21 - Han, Xueyuan, et al. "{SIGL}: Securing Software Installations Through Deep Graph Learning." *30th USENIX Security Symposium (USENIX Security 21)*. 2021.

Barr-Smith ’21 - Barr-Smith, Frederick, et al. "Survivalism: Systematic analysis of windows malware living-off-the-land." *2021 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2021.

Zeng ‘22 - Zeng, Jun, et al. "Shadewatcher: Recommendation-guided cyber threat analysis using system audit records." *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2022.

Goyal ’23 - Goyal, Akul, et al. "Sometimes, You Aren’t What You Do: Mimicry Attacks against Provenance Graph Host Intrusion Detection Systems." *30th ISOC Network and Distributed System Security Symposium (NDSS’23), San Diego, CA, USA*. 2023.

Colonial – Easterly, Jen “The Attack on Colonial Pipeline: What We’ve Learned & What We’ve Done over the Past Two Years: CISA.” Cybersecurity and Infrastructure Security Agency CISA, 8 Aug. 2023, www.cisa.gov/news-events/news/attack-colonial-pipeline-what-weve-learned-what-weve-done-over-past-two-years.

SolarWinds - “The Solarwinds Cyber-Attack: What You Need to Know.” *CIS*, 9 Nov. 2021, www.cisecurity.org/solarwinds.

References

McMahan ‘17 - McMahan, Brendan, et al. "Communication-efficient learning of deep networks from decentralized data." Artificial intelligence and statistics. Proceedings of Machine Learning Research, 2017.

Dang ’19 - Dang, Fan, et al. "Understanding fileless attacks on linux-based iot devices with honeycloud." Proceedings of Annual International Conference on Mobile Systems, Applications, and Services. 2019.

Hassan ’19 - Hassan, Wajih Ul, et al. "Nodoze: Combatting threat alert fatigue with automated provenance triage." Network and Distributed Systems Security Symposium. 2019.

Nguyen ‘19 – Nguyen, Thien Duc, et al. “DÏoT: A Federated Self-learning Anomaly Detection System for IoT.” IEEE 39th International Conference on Distributed Computing Systems. 2019.

Wang ’20 - Wang, Qi, et al. "You Are What You Do: Hunting Stealthy Malware via Data Provenance Analysis." Network and Distributed Systems Security Symposium. 2020.

Cosson ‘21 – Cossan, Adrien, et al. “Sentinel: A Robust Intrusion Detection System for IoT Networks Using Kernel-Level System Information.” Proceedings of the International Conference on Internet-of-Things Design and Implementation. 2021.

Han ‘21 - Han, Xueyuan, et al. "SIGL: Securing Software Installations Through Deep Graph Learning." USENIX Security Symposium. 2021.

Mukherjee ‘23 – Mukherjee, Kunal, et al. “Evading Provenance-Based ML Detectors with Adversarial System Actions." USENIX Security Symposium. 2023.

Rieger ‘23 – Rieger, Phillip, et al. “ARGUS: Context-Based Detection of Stealthy IoT Infiltration Attacks.” USENIX Security Symposium. 2023.

FritzFrog – David, Ori. “Frog4Shell — FritzFrog Botnet Adds One-Days to Its Arsenal“. Akamai. 2024. https://www.akamai.com/blog/security-research/fritzfrog-botnet-new-capabilities-log4shell.