Evading Provenance-Based ML Detectors with Adversarial System Actions

Kunal Mukherjee, Joshua Wiedemeier, Tianhao Wang, James Wei,
Feng Chen, Muhyun Kim, Murat Kantarcioglu, and Kangkook Jee

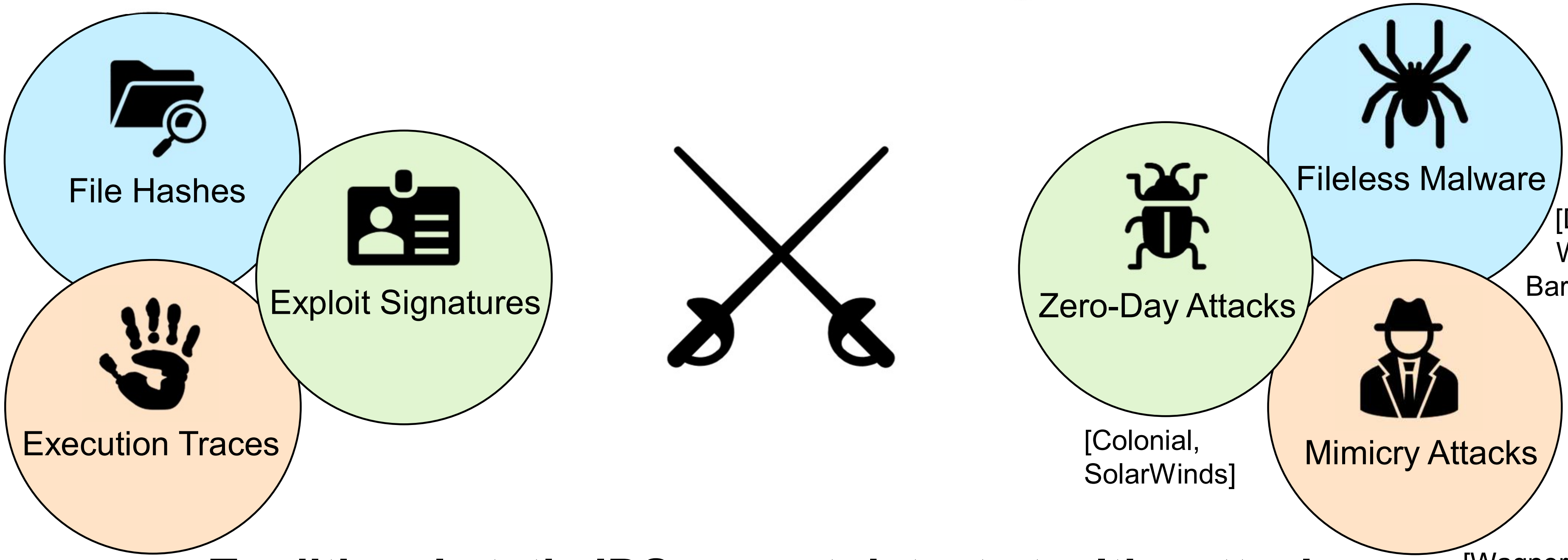
Department of Computer Science, The University of Texas at Dallas

USENIX Security 2023

Stealthy Attacks against Static Host Defenses

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

Adversaries evade detection with **stealthy techniques**



File Hashes

Exploit Signatures

Execution Traces

Fileless Malware

Zero-Day Attacks

Mimicry Attacks

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

[Colonial, SolarWinds]

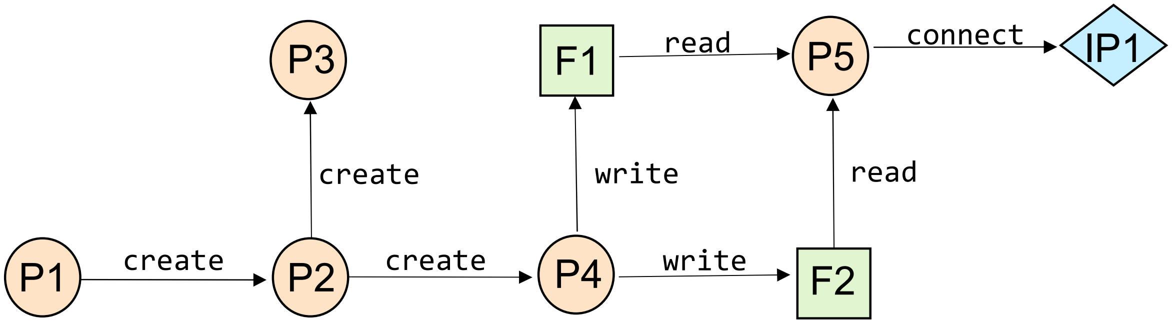
[Wagner & Soto '02, Tan & Maxion '03]

Traditional static IDS cannot detect stealthy attacks

Dynamic Defense against Stealthy Attacks

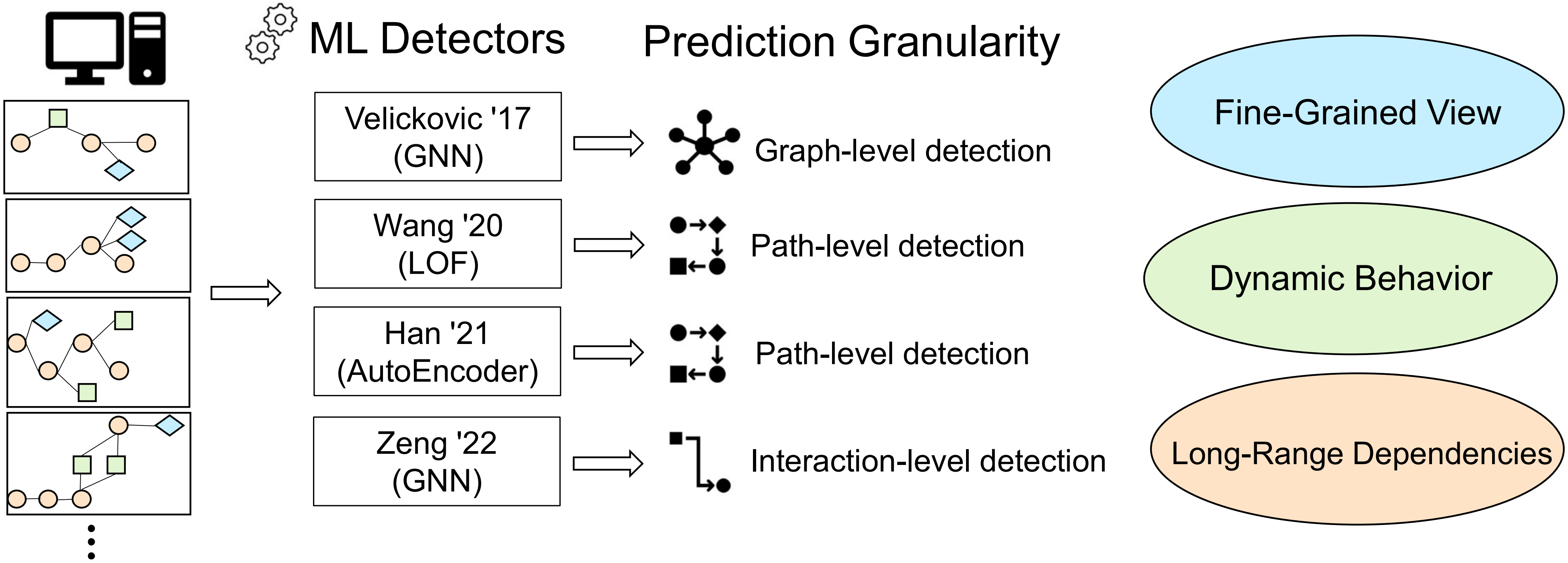
- **System Provenance** championed as a *host-based* dynamic defense
  - Influential works [Hassan '19, Wang '20, Han '21]

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



How can system Provenance help detect stealthy attacks?

Provenance-Based IDS



ML Detectors

Prediction Granularity

Velickovic '17 (GNN) → Graph-level detection

Wang '20 (LOF) → Path-level detection

Han '21 (AutoEncoder) → Path-level detection

Zeng '22 (GNN) → Interaction-level detection

Fine-Grained View

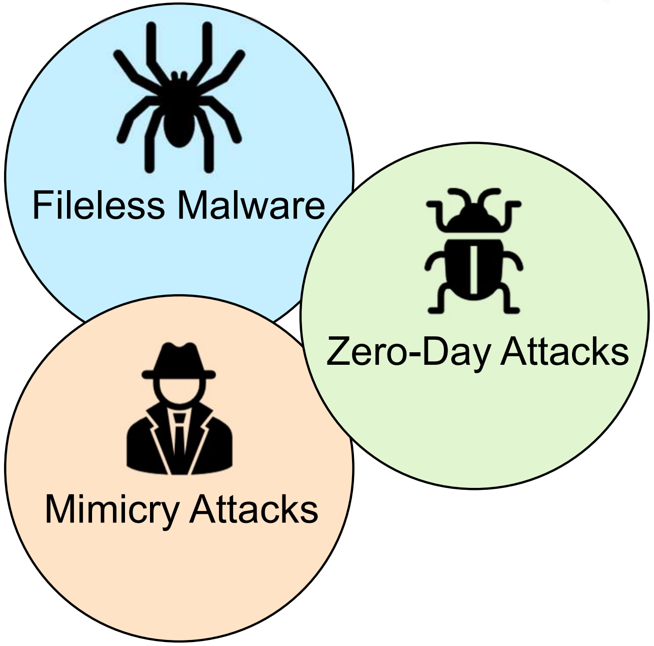
Dynamic Behavior

Long-Range Dependencies

Why are Provenance-based IDS gaining popularity?

Popularity of Provenance-Based IDS

Provenance captures **runtime behaviors**

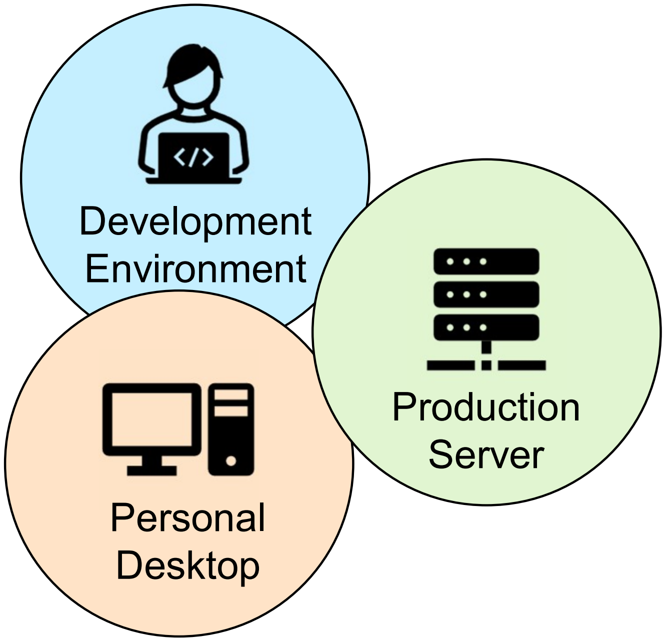


Fileless Malware

Zero-Day Attacks

Mimicry Attacks

ML models are **fine-tuned** for different environments

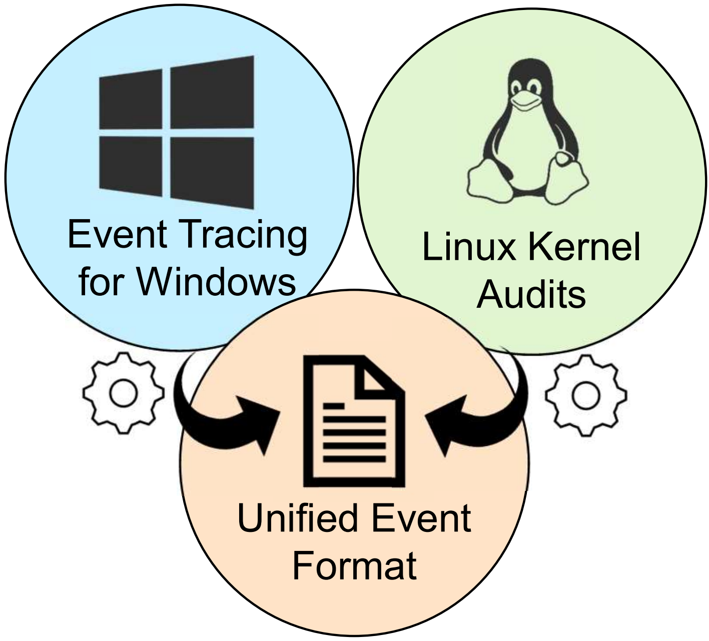


Development Environment

Production Server

Personal Desktop

Event collection frameworks provide **platform independence**



Event Tracing for Windows

Linux Kernel Audits

Unified Event Format

However, Provenance-based IDS are not yet mature.

Primary Roadblock to Provenance-Based IDS Adoption

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

Robustness against dedicated adversaries has **not been verified**

Adversarial validation is an established way to **prove robustness**

Adversarial Validation in Provenance-Based IDS

Generic adversarial techniques fail

- Heterogenous graphs with node/edge attributes

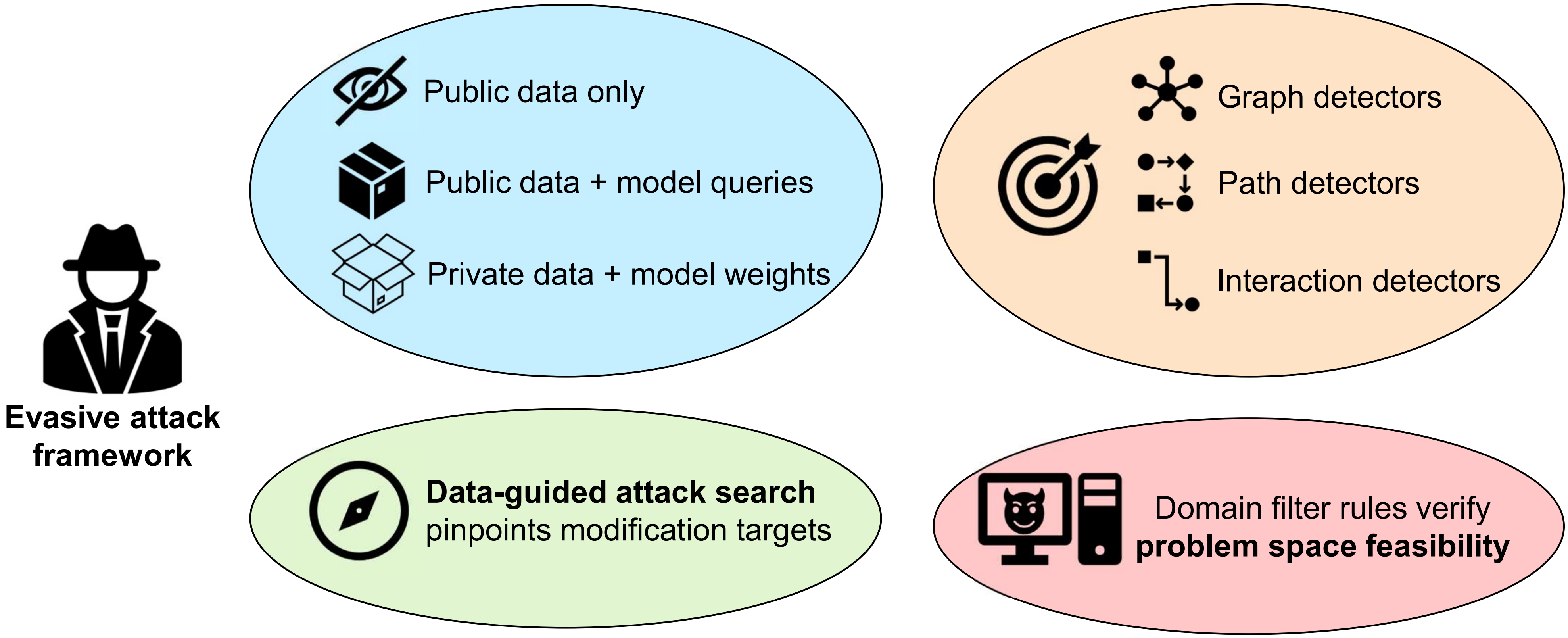
Problem space feasibility is critical for validation

- Only real-world attacks can invalidate defenses

Provenance mimicry attacks exist [Goyal '23], *however*

- **Require adding >15,000 events**
- **Require knowledge of the defense model architecture**
- **Unlikely to be effective against event-level detectors**

Contributions



Evasive attack framework

- Public data only
- Public data + model queries
- Private data + model weights

- Graph detectors
- Path detectors
- Interaction detectors

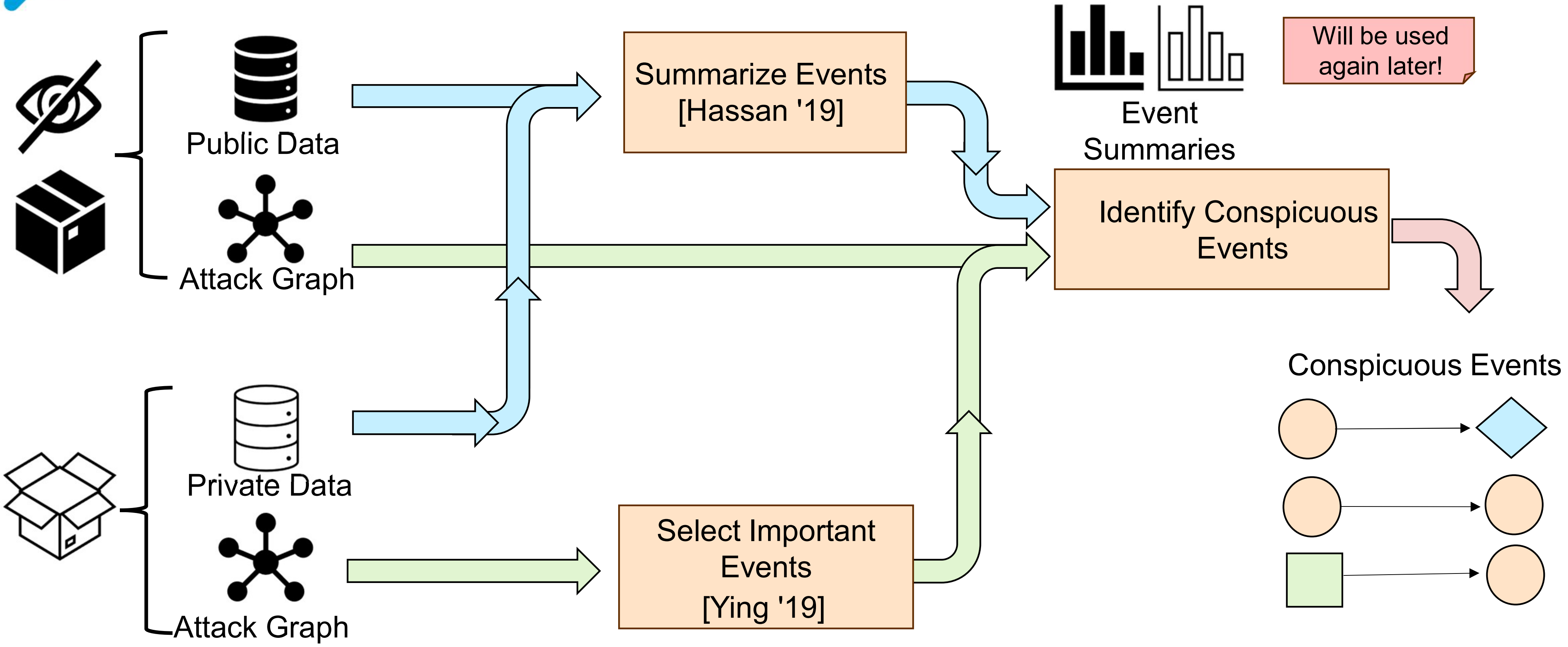
Data-guided attack search pinpoints modification targets

Domain filter rules verify **problem space feasibility**

ProvNinja: Evasive Attack Framework

- Identify Conspicuous Events
- Replace with Common Events
- Camouflage Processes
- Realize the Evasion

Identify Conspicuous Events



Public Data

Attack Graph

Summarize Events [Hassan '19]

Event Summaries

Will be used again later!

Identify Conspicuous Events

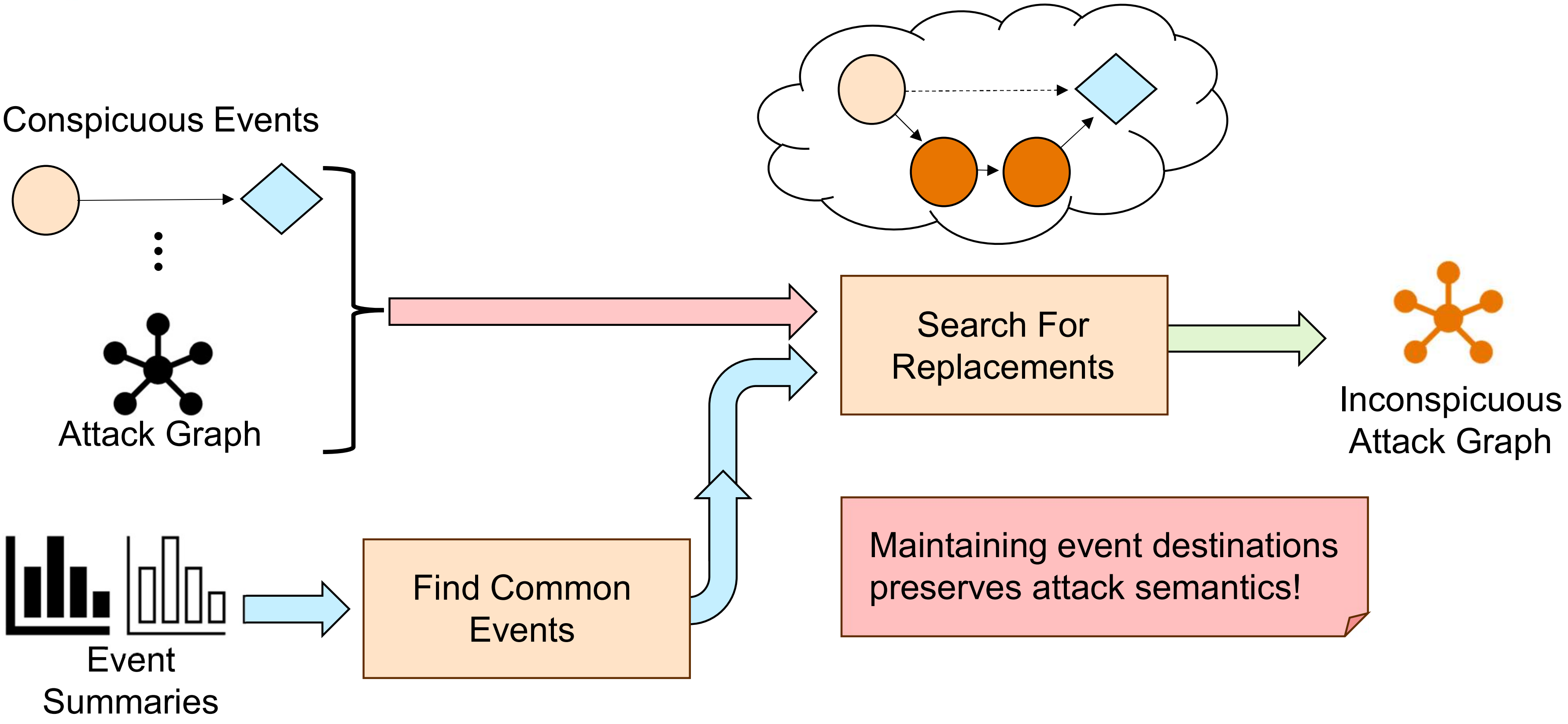
Conspicuous Events

Private Data

Attack Graph

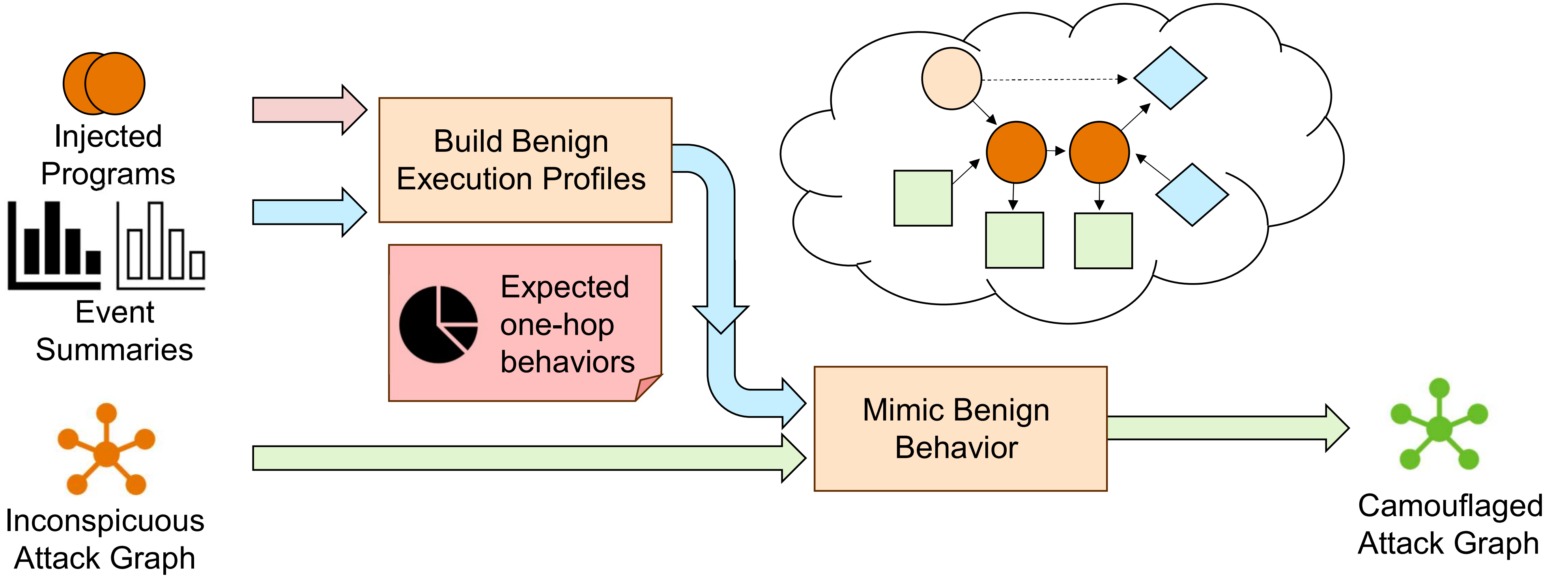
Select Important Events [Ying '19]

Replace with Common Events



Maintaining event destinations preserves attack semantics!

Camouflage Processes



Injected Programs

Event Summaries

Inconspicuous Attack Graph

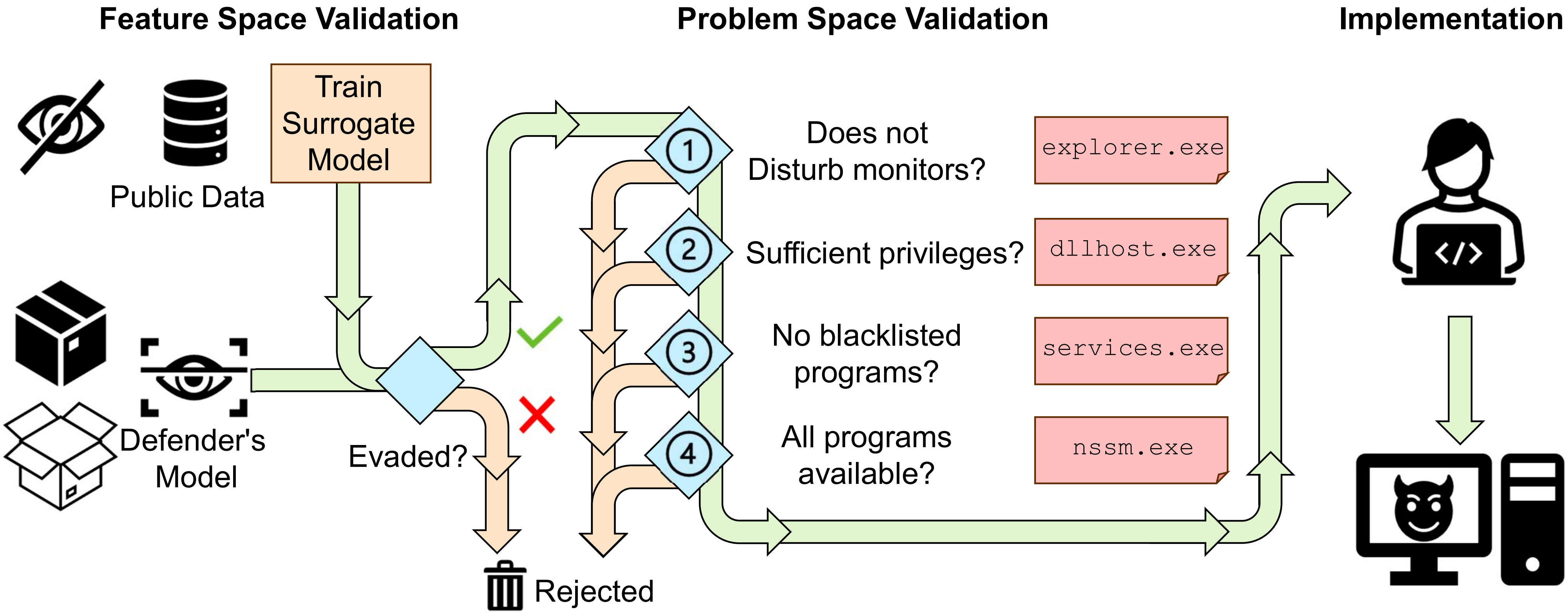
Build Benign Execution Profiles

Expected one-hop behaviors

Mimic Benign Behavior

Camouflaged Attack Graph

Realize the Evasion



Evaluation

- Datasets
- Experimental Setup
- Evasion Evaluation
- Realizability Evaluation

Datasets

Benign Datasets

DARPA (public)	In-House (private)
Scripted	Real Users
8 Hosts	86 Hosts
12 Days	13 Months

Malicious Datasets

Enterprise	**1,779** Graphs
Supply Chain	**1,091** Graphs
Fileless Malware [Barr-Smith '21]	**1,206** Graphs

Experimental Setup

Threat Models

- Blind: Public data only
- Black-box: Public data + model queries
- White-box: Private data + model weights

Provenance-based IDS

- [Veličković '17]
- [Wang '20, Han '21]
- [Zeng '22]

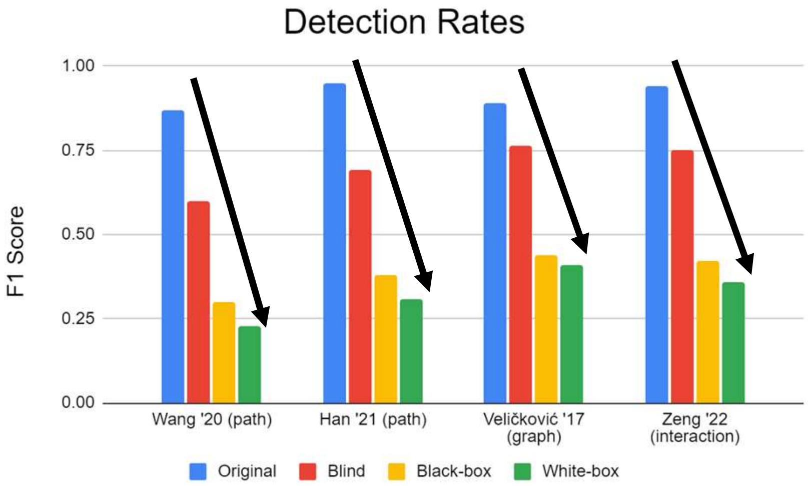
Dataset Allocation

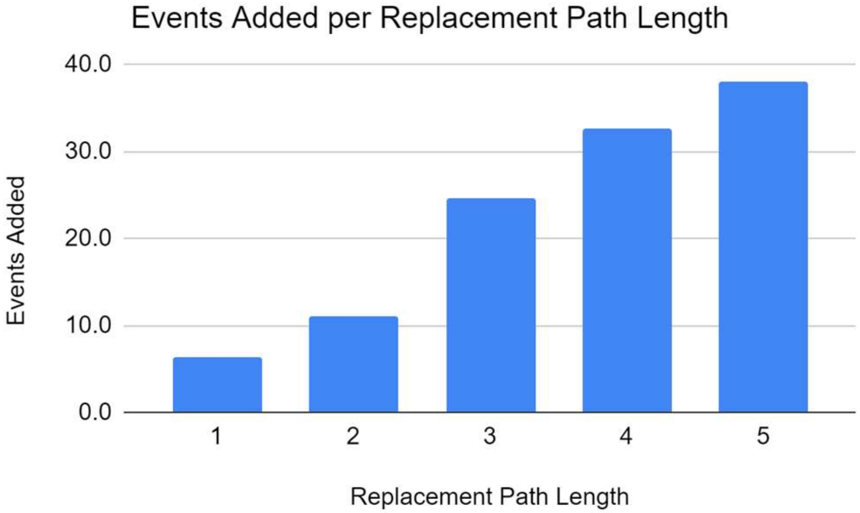
Public

DARPA

Private

Evasion Evaluation



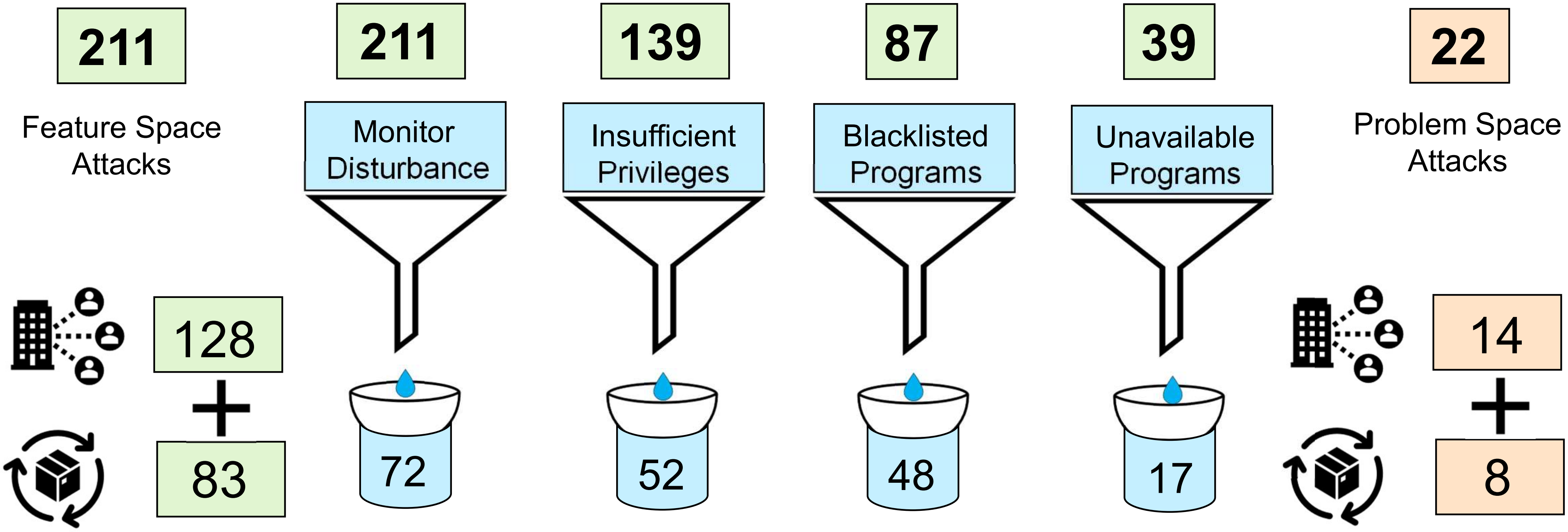


Reduces detection rates against SOTA Provenance-based IDS

Scales to threat model

Each replacement adds fewer than 40 events

Attack Realizability



211

Feature Space Attacks

128 + 83

211

Monitor Disturbance

72

139

Insufficient Privileges

52

87

Blacklisted Programs

48

39

Unavailable Programs

17

22

Problem Space Attacks

14 + 8

Conclusion

ProvNinja **systematically challenges** Provenance-based IDS

57%
Average detection rate reduction

<150
Average events added per attack

Transfers behavioral insights across environments

Supports adversarial testing and verification

Inspiring the development of **robust** IDS with **realistic** adversarial examples

THANK YOU

Please forward any questions, comments and future collaboration opportunities to
kxm180046@utdallas.edu

Scan the QR code to access the paper

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 &amp; 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.