Graph-based Network Intrusion Detection: The Data Imperative

Saikat DEY & Wu FENG
{dsaikat, wfeng}@vt.edu

S. Dey and W. Feng, "Graph-based Network Intrusion Detection: The Data Imperative,"
19th Workshop on Active Internet Measurements (AIMS), La Jolla, CA, February 2026.

VIRGINIA TECH

SyNeRG

synergy.cs.vt.edu

Graph-based Network Intrusion Detection (NID)

Why a graph-based approach to NID?

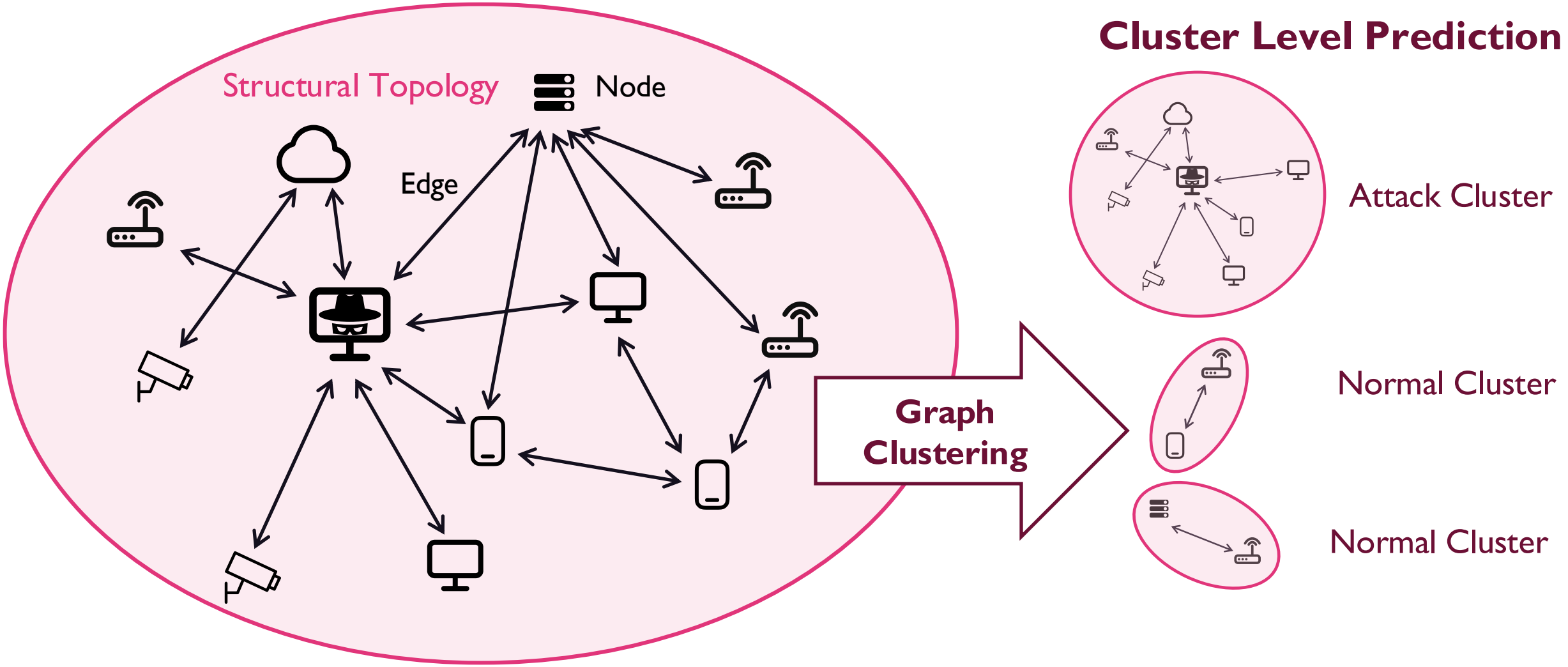
- Offers a natural representation of network flows.
- Detects anomalies via interaction patterns and exposes insidious threats that look benign in isolation.
- Makes attackers masking of their structural communication significantly more difficult.
  - ✓ Outcome: Robust method for NID

The Data Imperative

- Approaches like *graph clustering* need:
  - ✓ Structural topology
- Approaches like *graph neural networks* (GNNs) additionally need:
  - ✓ Large graph dataset
  - ✓ Meta-level information
    - ***temporal information***
    - ***rich attributes (at nodes and edges)***
  - ✓ Ground truth

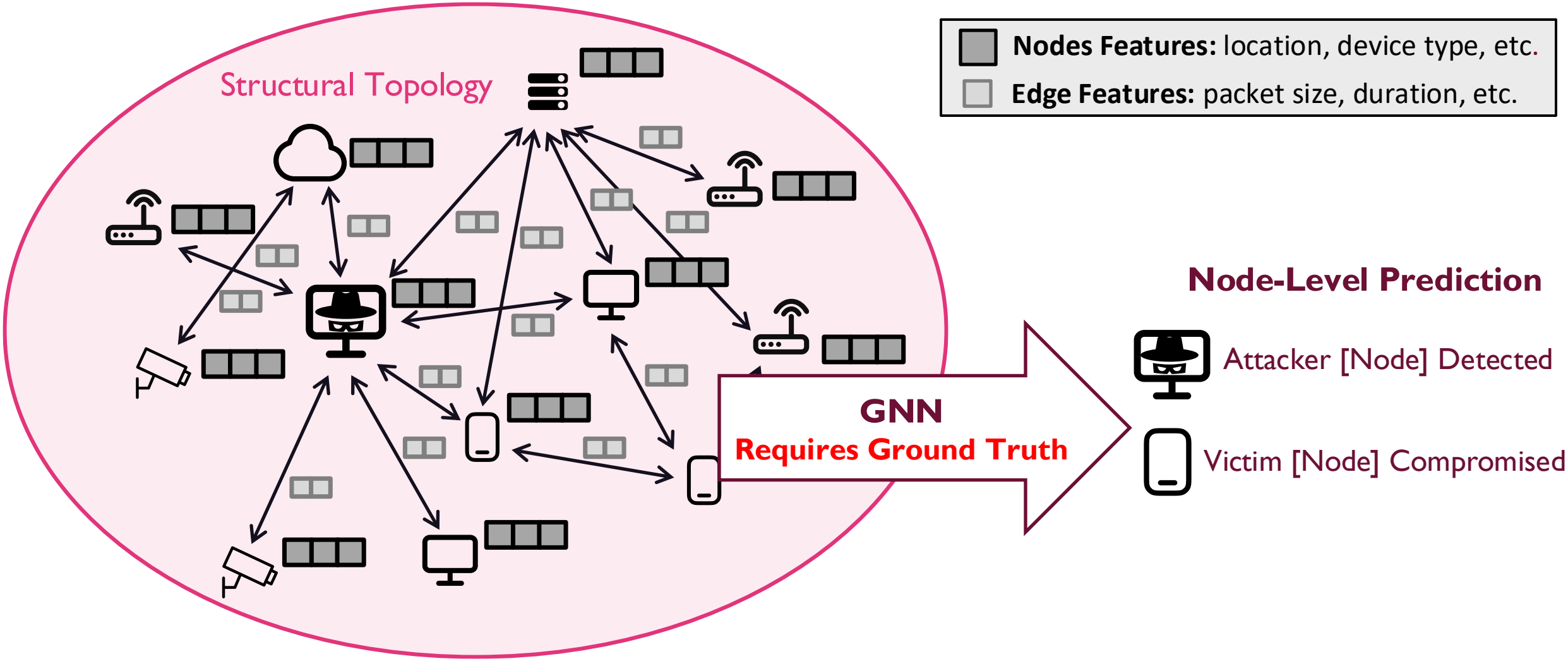
S. Dey and W. Feng, "Graph-based Network Intrusion Detection: The Data Imperative," 19th Workshop on Active Internet Measurements (AIMS), La Jolla, CA, February 2026.

Approach: Graph Clustering



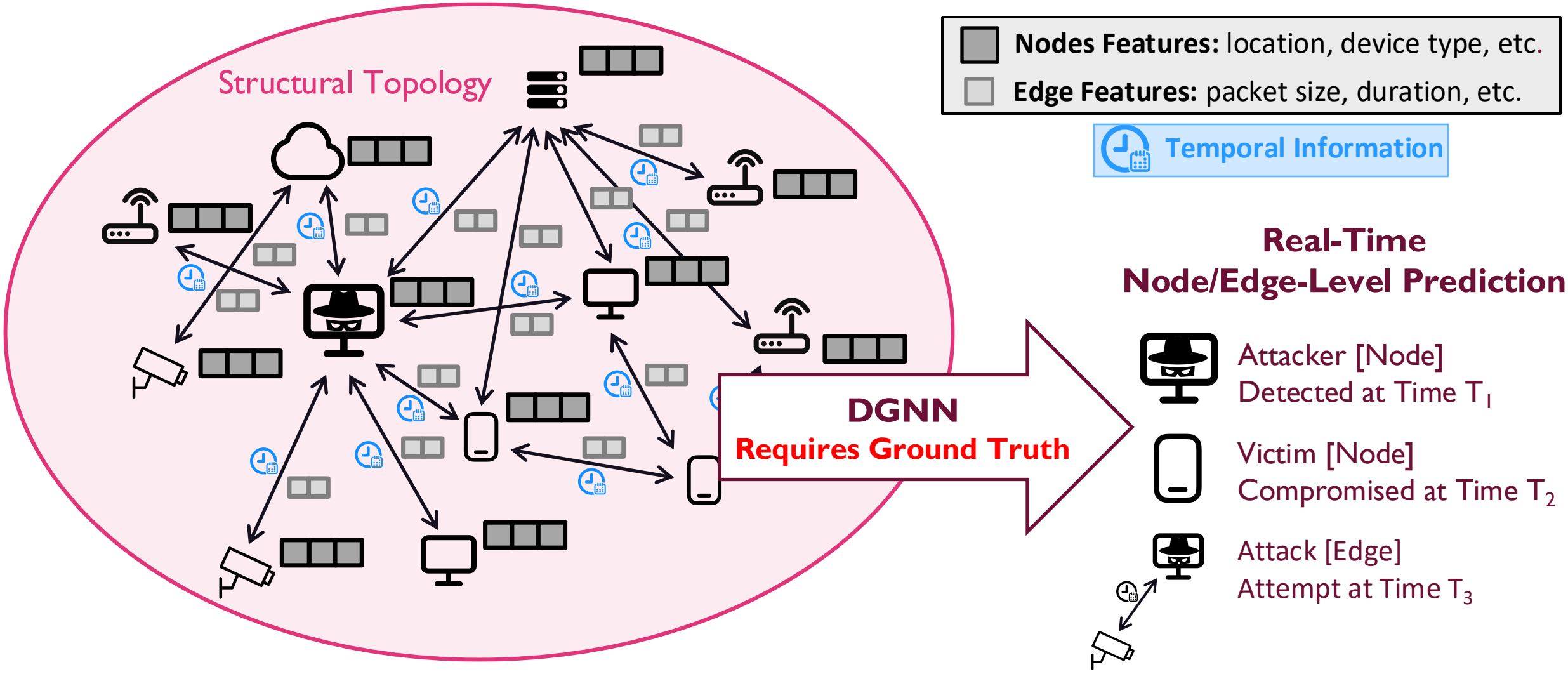
S. Dey and W. Feng, "Graph-based Network Intrusion Detection: The Data Imperative," 19th Workshop on Active Internet Measurements (AIMS), La Jolla, CA, February 2026.

Approach: Static GNNs



S. Dey and W. Feng, "Graph-based Network Intrusion Detection: The Data Imperative," 19th Workshop on Active Internet Measurements (AIMS), La Jolla, CA, February 2026.

Approach: Dynamic GNNs



S. Dey and W. Feng, "Graph-based Network Intrusion Detection: The Data Imperative," 19th Workshop on Active Internet Measurements (AIMS), La Jolla, CA, February 2026.

Data Needs for Graph-based Network Intrusion Detection (NID)

Question: What does good data look like for graph-based NID?

Data Needs for Graph-based Network Intrusion Detection (NID)

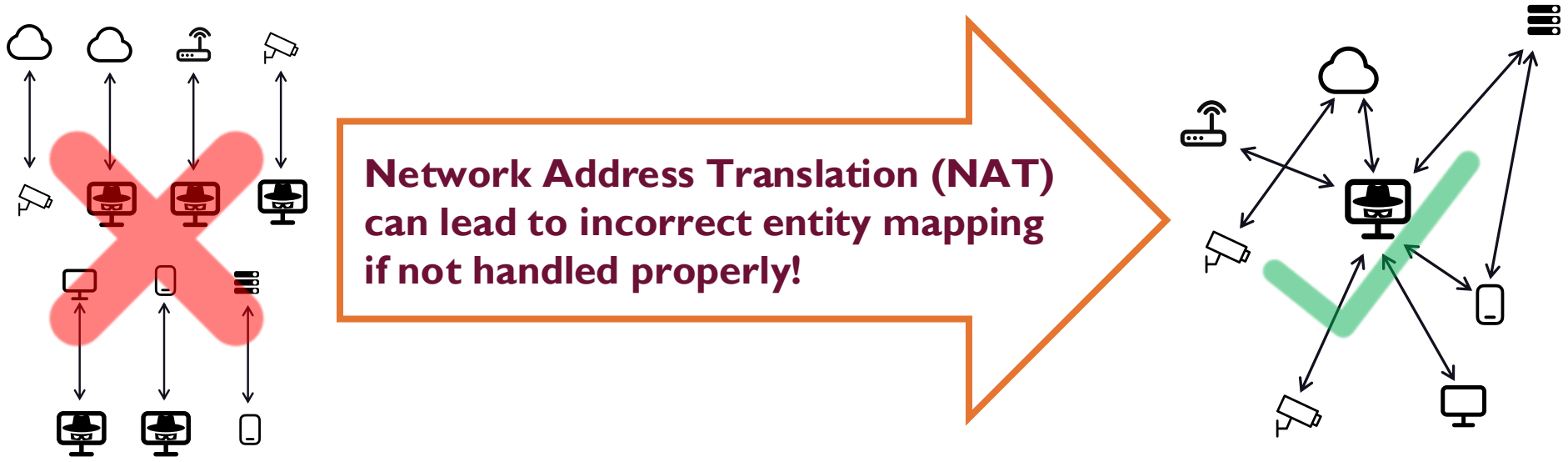
Question: What does good data look like for graph-based NID?

- **Availability of Graphs at a Large-scale**
  - GNNs are particularly data hungry and require both large *number* of graphs and larger *size* of graphs.
    - Number of graphs in the order of *100s*.
    - Number of nodes in the order of *millions*.
    - Number of edges in the order of *10s of millions*.

Data Needs for Graph-based Network Intrusion Detection (NID)

Question: What does good data look like for graph-based NID?

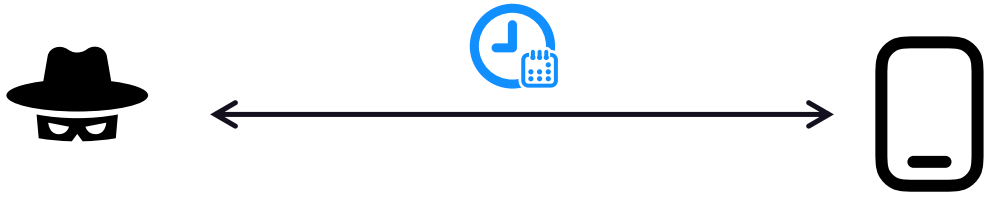
- **Availability of Graphs at a Large-scale**
  - GNNs are particularly data hungry and require both large *number* of graphs and larger *size* of graphs.
    - Number of graphs in the order of *100s*.
    - Number of nodes in the order of *millions*.
    - Number of edges in the order of *10s of millions*.
- **Availability of Network's (Graph's) Structural Topology**
  - All network entities (nodes) and their connections (edges) to other network entities should be known (e.g., PCs, Routers, IoTs, Cloud, DNS, etc.).



Data Needs for Graph-based Network Intrusion Detection (NID)

Question: What does good data look like for graph-based NID?

- **Availability of Graphs at a Large-scale**
  - GNNs are particularly data hungry and require both large *number* of graphs and larger *size* of graphs.
    - Number of graphs in the order of *100s*.
    - Number of nodes in the order of *millions*.
    - Number of edges in the order of *10s of millions*.
- **Availability of Network's (Graph's) Structural Topology**
  - All network entities (nodes) and their connections (edges) to other network entities should be known (e.g., PCs, Routers, IoTs, Cloud, DNS, etc.).
- **Availability of Temporal Information**
  - All connections (edges) should have an associated timestamp.

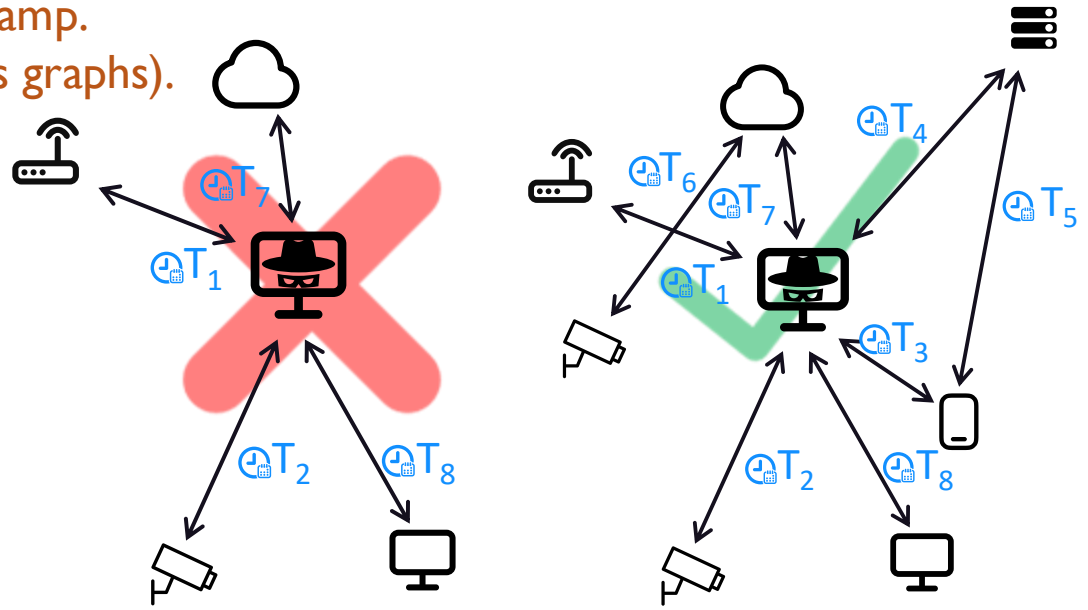


Timestamp = 02-22-2026; 9:38 PM

Data Needs for Graph-based Network Intrusion Detection (NID)

Question: What does good data look like for graph-based NID?

- **Availability of Graphs at a Large-scale**
  - GNNs are particularly data hungry and require both large *number* of graphs and larger *size* of graphs.
- **Availability of Network's (Graph's) Structural Topology**
  - All network entities (nodes) and their connections (edges) to other network entities should be known (e.g., PCs, Routers, IoTs, Cloud, DNS, etc.).
- **Availability of Temporal Information**
  - All connections (edges) should have an associated timestamp.
  - There should be no break in the data stream (continuous graphs).

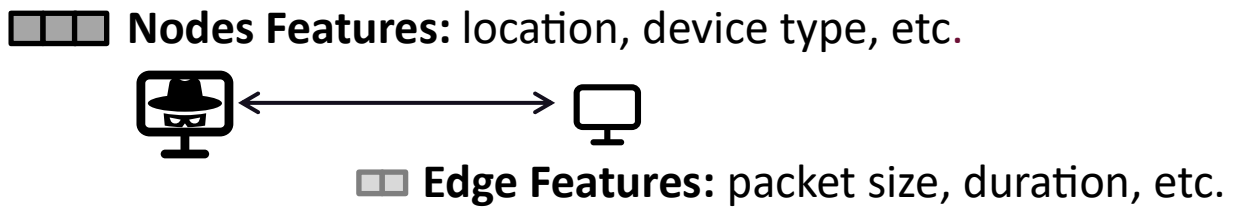


S. Dey and W. Feng, "Graph-based Network Intrusion Detection: The Data Imperative," 19th Workshop on Active Internet Measurements (AIMS), La Jolla, CA, February 2026.

Data Needs for Graph-based Network Intrusion Detection (NID)

Question: What does good data look like for graph-based NID?

- **Availability of Graphs at a Large-scale**
  - GNNs are particularly data hungry and require both large *number* of graphs and larger *size* of graphs.
    - Number of graphs in the order of *100s*.
    - Number of nodes in the order of *millions*.
    - Number of edges in the order of *10s of millions*.
- **Availability of Network's (Graph's) Structural Topology**
  - All network entities (nodes) and their connections (edges) to other network entities should be known (e.g., PCs, Routers, IoTs, Cloud, DNS, etc.).
- **Availability of Temporal Information**
  - All connections (edges) should have an associated timestamp.
  - There should be no break in the data stream (continuous graphs).
- **Availability of Node and Edge Features**
  - As much information regarding entities (nodes) and connections (edges) should be available.



Data Needs for Graph-based Network Intrusion Detection (NID)

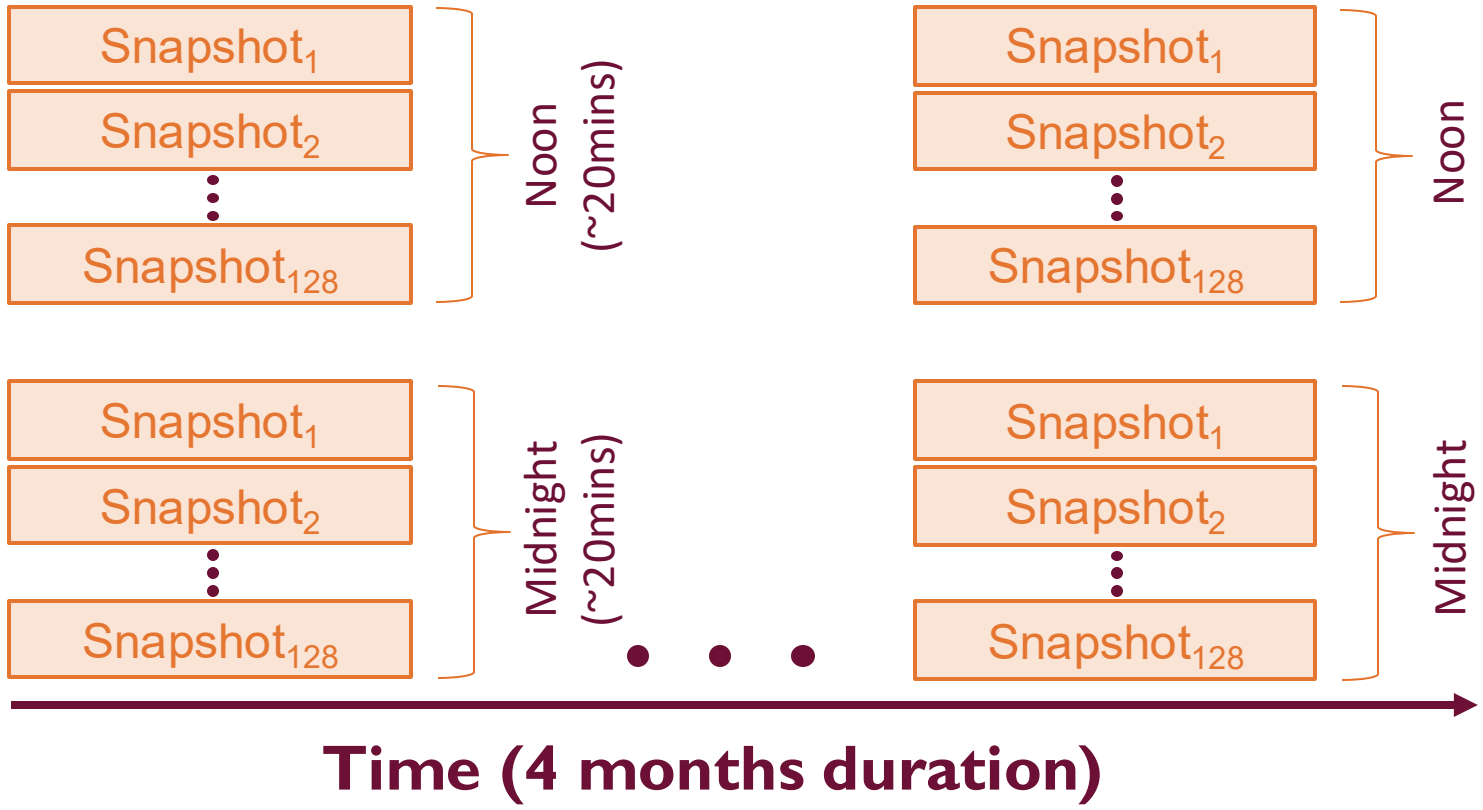
Question: What does good data look like for graph-based NID?

- **Availability of Graphs at a Large-scale**
  - GNNs are particularly data hungry and require both large *number* of graphs and larger *size* of graphs.
    - Number of graphs in the order of *100s*.
    - Number of nodes in the order of *millions*.
    - Number of edges in the order of *10s of millions*.
- **Availability of Network's (Graph's) Structural Topology**
  - All network entities (nodes) and their connections (edges) to other network entities should be known (e.g., PCs, Routers, IoTs, Cloud, DNS, etc.).
- **Availability of Temporal Information**
  - All connections (edges) should have an associated timestamp.
  - There should be no break in the data stream (continuous graphs).
- **Availability of Node and Edge Features**
  - As much information regarding entities (nodes) and connections (edges) should be available.
- **Availability of Ground Truth**
  - Very important both from training and testing point of view.
  - Attacker (node), attack timeline (edges), and victims (nodes) are important.

S. Dey and W. Feng, "Graph-based Network Intrusion Detection: The Data Imperative," 19th Workshop on Active Internet Measurements (AIMS), La Jolla, CA, February 2026.

Network Telescope (CAIDA) graphs

MIT launched its latest Graph Challenge in 2024 –
The **Anonymized Network Sensing Graph Challenge**, which uses the Network Telescope (CAIDA) dataset. Link: https://graphchallenge.mit.edu/challenges



- Each snapshot can be viewed as an individual graph snapshot.
- Combining [$Snapshot_1 \ldots Snapshot_{128}$] creates a full static graph.
- There are 161 static graphs, with 128 snapshots in each, totaling 20,608 graph snapshots.
- The data is enriched using the GreyNoise honeyfarm database, which provides labels for certain IPs.

Overview: Graph-based Network Intrusion Detection (NID)

Required Components of the Graph Dataset by NID Approaches

Overview: Graph-based Network Intrusion Detection (NID)

Required Components of the Graph Dataset by NID Approaches

NID Approaches	No. of Graphs Required for Training	Structural Topology	Temporal Information	Node/Edge Attributes	Ground Truth
Graph Clustering [1]	N/A	✓	✗	✗	✗
Static GNNs	1 graph; >1M nodes/graph	✓	✗	✓	✓
Dynamic GNNs [2]	100+ graphs; >1M nodes/graph	✓	✓	✓	✓

Overview: Graph-based Network Intrusion Detection (NID)

Required Components of the Graph Dataset by NID Approaches

NID Approaches	No. of Graphs Required for Training	Structural Topology	Temporal Information	Node/Edge Attributes	Ground Truth
Graph Clustering [1]	N/A	✓	✗	✗	✗
Static GNNs	1 graph; >1M nodes/graph	✓	✗	✓	✓
Dynamic GNNs [2]	100+ graphs; >1M nodes/graph	✓	✓	✓	✓

Available Components of the Graph Dataset in NID Datasets

Overview: Graph-based Network Intrusion Detection (NID)

Required Components of the Graph Dataset by NID Approaches

NID Approaches	No. of Graphs Required for Training	Structural Topology	Temporal Information	Node/Edge Attributes	Ground Truth
Graph Clustering [1]	N/A	✓	✗	✗	✗
Static GNNs	1 graph; >1M nodes/graph	✓	✗	✓	✓
Dynamic GNNs [2]	100+ graphs; >1M nodes/graph	✓	✓	✓	✓

Available Components of the Graph Dataset in NID Datasets

NID Datasets	No. of Graphs Available	Structural Topology	Temporal Information	Node/Edge Attributes	Ground Truth
Network Telescope (CAIDA) graphs [5]	161 graphs; ~8M nodes/graph	✓	⚠ (Snapshots Only)	✗	⚠ (Partial)
G2A2 [3]	12 graphs; ~0.7M nodes/graph	✓	✓	✓	✓
P-core [4]	2 graphs; ~4M nodes/graph	✓	✓	✓	✓

[1] Dey, Jha, Wanye, Feng, *On the Landscape of Graph Clustering at Scale*. IEEE IPDPS Workshops, 2025.
[2] Dey, Gardner, Lang, Feng, *STAGS: A Graph-Sampling Approach for GNN-based Network Anomaly Detection*. IFIP Networking, 2025.
[3] Dey, Jha, Feng, *G2A2: An Automated Graph Generator with Attributes and Anomalies*. ACM Computing Frontiers (CF), 2024.
[4] Buchanan et al., *On Generating and Labeling Network Traffic with Realistic, Self-Propagating Malware*. SIAM SDM AI/ML for Cybersecurity Workshop, 2021.
[5] Jananthan et al., *Anonymized Network Sensing Graph Challenge Dataset*. CAIDA, 2022.