

When Network Topology Supports Graph-Based Intrusion Detection: A Pre-Deployment Validity Framework

Abdulhadi Albluwi

`Abdelhady_pgs@fci.capu.edu.eg`

Helwan University

Mohamed I. Marie

Helwan University

Helal A. Suleiman

Helwan University

Research Article

Keywords: graph structural differentiability, prominence isolation, threshold collapse, pre-deployment diagnostic, graph-based intrusion detection, topology-conditioned reasoning

Posted Date: September 3rd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10707774/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Graph-based intrusion detection can require substantial effort in graph construction, model development, and operational deployment, yet its usefulness depends on whether the underlying network topology actually contains a discriminative structural signal. Applying such methods to an unsuitable environment can therefore consume resources without providing reliable security value. This study addresses a prior deployment question: **when does network topology actually support graph-based intrusion detection?**

We distinguish two complementary dimensions of structural validity: **Graph Structural Differentiability (GSD)**, which characterizes whether meaningful structural differentiation exists, and **Prominence Isolation (PI)**, which assesses whether the structurally prominent region is sufficiently isolated to support attack-selective reasoning. Threshold Collapse is treated as a threshold-failure mechanism, separating homogeneous collapse from recoverable population-expansion collapse using label-free topology diagnostics.

Evaluation across five heterogeneous benchmark environments reveals a clear dissociation between structural differentiation and attack-selective prominence. In a matched full-scale comparison, NF-CSE-CIC-IDS2018-v2 shows higher aggregate structural inequality than NF-ToN-IoT-v2 (Gini = 0.812 vs. 0.271) yet yields Hybrid Precision = 0.000 under gap-based selection. By contrast, NF-ToN-IoT-v2 achieves Precision = 0.909 and Lift $\approx 677\times$ after topology-adaptive recovery. Production validation over 23.8 million HTTPS requests further shows that structural validity is temporally sensitive as observation windows expand.

These findings show that structural differentiation alone is insufficient to justify graph-based intrusion reasoning. The proposed pre-deployment framework distinguishes structurally suitable, recoverable, infrastructure-dominated, and structurally degenerate environments before committing resources to graph-based intrusion detection deployment.

1. Introduction

Cyberattacks impose substantial economic, operational, and reputational costs on organizations and critical infrastructures. Global cybercrime damage has been estimated at USD 10.5 trillion annually in 2025 [1], while the average cost of a data breach remains in the multi-million-dollar range [2]. Intrusion detection is therefore not only a model-optimization problem but also a deployment decision: organizations must determine whether a detection architecture is appropriate for the environment before committing resources to graph construction, model training, computational infrastructure, and analyst workflows.

Graph-based intrusion detection has emerged as a promising approach for exploiting relationships that are not represented by isolated flow records. By modeling communication as a graph, these methods use structural properties such as connectivity, centrality, and entity prominence to identify

communication patterns associated with malicious activity [3, 4]. Recent work has extended this paradigm through graph neural networks, temporal graph learning, and hybrid graph-behavioral architectures [5–7]. Despite this methodological progress, evaluation remains predominantly model-centric: a graph is constructed from a chosen environment, a detector is applied, and the resulting predictive performance is measured. The topology itself is generally treated as an available source of structural evidence rather than as a condition whose suitability for intrusion reasoning must first be established.

This assumption creates a prior validity problem. Structural heterogeneity does not necessarily imply security-relevant structural discrimination. High-connectivity entities may represent malicious scanning or coordinated attack activity, but they may also be legitimate servers, gateways, crawlers, or other infrastructure nodes. Conversely, apparent failure of a percentile-based structural detector may result either from a genuine absence of discriminative topology or from a thresholding artifact that obscures a small but strongly isolated structural tail. Graph-based detection failure can therefore originate from at least three different sources: the detection model, the thresholding mechanism, or the topology itself. Distinguishing these causes is essential because improving the model cannot recover structural evidence that the deployment topology does not provide.

To address this prior validity problem, we distinguish two properties that must not be conflated in graph-based intrusion reasoning: **structural differentiation** and **attack-selective prominence**. The first concerns whether the communication topology contains structurally distinguishable connectivity regimes; the second concerns whether the structurally prominent region represents a security-relevant signal rather than legitimate high-connectivity infrastructure. We operationalize these properties through Graph Structural Differentiability (GSD) and Prominence Isolation (PI), respectively. GSD characterizes the existence and strength of structural differentiation from unlabeled traffic, whereas PI evaluates whether structural prominence is sufficiently isolated to support attack-selective reasoning. The central premise examined in this study is that these properties are empirically dissociable: a topology may exhibit substantial structural differentiation while providing little or no useful structural intrusion signal when prominence is infrastructure-dominated.

This validity distinction also requires separating failure of the topology from failure of the thresholding mechanism. We therefore characterize **Threshold Collapse**, in which conventional percentile-based thresholds cease to provide meaningful structural separation, and distinguish two topologically different forms. **Homogeneous collapse** reflects a genuine absence of recoverable structural differentiation and therefore indicates that structural reasoning is unsuitable. **Population-expansion collapse**, by contrast, occurs when a large peripheral population causes percentile thresholds to degenerate even though an isolated structural tail remains present; in this regime, topology-adaptive gap-based thresholding can recover the latent structural signal. Threshold Collapse is thus treated not as an independent validity criterion, but as a mechanism for determining whether apparent structural failure is intrinsic to the topology or recoverable through threshold adaptation.

The empirical analysis is designed to test this validity distinction across heterogeneous network environments rather than merely to compare detection performance. Five benchmark environments spanning enterprise, IoT, and laboratory topologies are used to expose qualitatively different structural regimes. The most informative contrast is between NF-ToN-IoT-v2 and NF-CSE-CIC-IDS2018-v2. Although NF-CSE-CIC-IDS2018-v2 exhibits substantially greater aggregate structural inequality, its structurally prominent region is dominated by benign infrastructure and yields no attack-selective utility under gap-based structural selection. In contrast, NF-ToN-IoT-v2 exhibits lower aggregate inequality and percentile-threshold collapse, yet a strongly isolated prominent tail remains recoverable and achieves Hybrid Precision = 0.909 and Lift $\approx 677\times$. This contrast provides direct empirical evidence that structural differentiation and attack-selective prominence are distinct properties, and that graph-based structural reasoning should not be considered valid solely because a topology exhibits substantial structural heterogeneity.

The contributions of this study are threefold:

1. **Topology-conditioned structural validity.** We formulate the validity of graph-based intrusion reasoning as a pre-deployment topology problem and operationalize it through two complementary conditions: Graph Structural Differentiability (GSD), which characterizes whether meaningful structural differentiation exists, and Prominence Isolation (PI), which characterizes whether structurally prominent entities provide an attack-selective rather than infrastructure-dominated signal.
2. **Empirical GSD–PI dissociation.** We demonstrate that structural differentiation and attack-selective prominence can diverge across heterogeneous network environments. In a methodologically matched full-scale comparison, NF-CSE-CIC-IDS2018-v2 exhibits substantially greater aggregate structural inequality than NF-ToN-IoT-v2 (Gini = 0.812 versus 0.271), yet yields Hybrid Precision = 0.000 under gap-based structural selection, whereas NF-ToN-IoT-v2 achieves Precision = 0.909 and Lift $\approx 677\times$. This result establishes that aggregate structural inequality alone is insufficient to determine structural detection validity.
3. **Operational diagnosis and boundary validation.** We translate the GSD–PI distinction into a pre-deployment diagnostic procedure that differentiates irrecoverable homogeneous collapse from recoverable population-expansion collapse, applies topology-adaptive thresholding where appropriate, evaluates structural consistency across threshold-based, supervised, and unsupervised behavioral paradigms, and examines temporal validity under 23.8 million production HTTPS requests.

Building on a companion study establishing structural-behavioral convergence in a laboratory environment [8], this paper investigates the topological conditions that determine when and why such convergence holds – and identifies the boundary conditions under which it fails. Section 2 reviews related work. Section 3 presents the methodology. Section 4 describes the experimental setup. Section 5 presents empirical results, including the topology validity taxonomy and production web validation. Section 6 discusses implications and limitations. Section 7 concludes.

2. Related Work

2.1 Graph-Based Structural Approaches to Intrusion Detection

Graph-based methods have established a productive research direction by modeling network interactions as structured topologies rather than independent flow records. Zoubir and Missaoui [3] employed Graph Neural Networks to capture communication dependencies, demonstrating improvements in anomaly separation over flat feature-based methods. Wu et al. [5] extended this direction through temporal contrastive graph learning, achieving high detection accuracy and low false-alarm rates by capturing spatiotemporal traffic patterns. Govindarajan and Muzamal [9] combined graph-Transformer architectures with contrastive learning to achieve high precision across diverse attack types. Mwiga et al. [10] integrated graph convolutional networks with deep Q-networks, reporting strong classification performance with explainability analysis on benchmark datasets.

A characteristic shared across these contributions is that structural signals derived from the network graph are treated as inherently informative, with evaluation focused on model performance rather than on the topological conditions that make graph-derived signals discriminative. The prior validity of structural reasoning — whether the topology itself is heterogeneous enough to support meaningful structural separation — is not explicitly examined. The present paper addresses this examination gap directly.

A recent systematic review of dynamic temporal approaches to graph-based anomaly detection by Ares-Robledo et al. [11] identifies a consistent gap in the literature: while temporal GNN architectures have advanced substantially in their capacity to model evolving graph states, the validity of the structural signals they rely upon is rarely examined as a prior condition. Models are evaluated within fixed benchmark environments without assessing whether the underlying topology produces structurally discriminable entity populations in the first place. The present paper addresses precisely this prior condition, treating structural validity as a necessary prerequisite for graph-based reasoning rather than an implicit assumption.

2.2 Hybrid Graph-Behavioral Frameworks

Hybrid approaches combining structural graph analysis with behavioral anomaly detection have attracted growing attention. Farrukh et al. [6] developed XG-NID, a dual-modality system combining heterogeneous GNNs with large language models, demonstrating multi-modal detection capability. Belcastro et al. [12] combined knowledge graph reasoning with explainable AI techniques to improve decision transparency. Recent work on hybrid corroborative reasoning has examined the conditions under which structural and behavioral channels produce convergent high-confidence outputs without requiring labeled supervision [8].

A common characteristic of these hybrid architectures is that analytical integration occurs at the feature-level or through shared classifier pipelines, coupling detection performance to specific topological configurations and labeled training distributions. The question of whether structural and behavioral channels produce consistent outputs across heterogeneous behavioral paradigms — a form of structural robustness independent of the specific behavioral mechanism — has not been systematically examined in the literature.

2.3 Network Topology and Centrality-Based Detection

Centrality-based approaches to anomaly detection have a longer history in network security. Source out-degree, betweenness centrality, and PageRank-derived measures have been used to identify structurally prominent entities consistent with scanning, enumeration, and command-and-control behaviors. Gilliard et al. [7] demonstrated that knowledge graph reasoning can uncover latent attack paths through structural inference, noting that the integrity of the reasoning chain depends on the quality and completeness of the underlying graph. Alkahtani and Ahmad [13] developed autonomous knowledge graph construction systems integrating heterogeneous cybersecurity data sources.

These contributions implicitly assume that structurally prominent entities are meaningfully separable from the background population — an assumption that holds in topologically heterogeneous networks but fails in environments where out-degree distributions are near-uniform. The conditions under which centrality-based separation is reliable versus unreliable have not been formally characterized. GSD provides this characterization.

2.4 Validity and Generalizability of IDS Evaluation

A significant body of methodological work has examined the validity and generalizability of intrusion detection evaluation. Sommer and Paxson [14] argued that anomaly detection research faces fundamental challenges in moving from controlled laboratory evaluation to operational deployment, identifying the distributional mismatch between training and deployment environments as a central concern. Subsequent work has documented dataset-specific biases, label quality issues, and evaluation protocol weaknesses that limit the transferability of published results across environments.

Dilini et al. [4] conducted a systematic review of sixty graph-based anomaly detection studies, confirming that hybrid two-stage approaches account for 53% of reviewed studies while semantic explainability and cross-environment generalizability remain underexplored. Sarhan et al. [15] demonstrated zero-shot machine learning for threat detection through attribute-based class representations, addressing generalization beyond training categories without structural context.

The limitations of short-duration benchmark datasets for evaluating temporal aspects of detection validity have been underscored by the release of CESNET-TimeSeries24 [16], a dataset comprising 40 weeks of traffic from more than 275,000 active IP addresses in the CESNET3 ISP network. Koumar et al. [16] demonstrate that real-world network traffic exhibits sustained non-stationarity — including behavioral shifts, seasonal patterns, and entity population dynamics — that is fundamentally absent from

the static, short-duration captures characteristic of standard IDS benchmarks. This finding provides independent empirical grounding for the temporal validation conducted in § 5.9 of the present paper, where a 30-day production web environment reveals PI dynamics that are invisible under the static evaluation paradigm.

These contributions address generalizability at the model and distribution level. The present paper addresses a prior, topology-level validity question: whether the structural environment itself supports the type of reasoning the model is designed to perform. This distinction — between model generalizability and structural validity — has not been drawn explicitly in prior work. The present paper further extends this distinction to a temporal dimension by demonstrating, through production web validation (§ 5.9), that structural validity is not a fixed environmental property but a time-varying state conditioned on observation-window length, entity persistence patterns, and external traffic events — a form of temporal non-stationarity complementary to the dataset-level dynamics documented by Koumar et al. [16] and the modeling gaps identified by Ares-Robledo et al. [11].

2.5 Explainability in Security Reasoning

The integration of explainability into security systems has gained momentum in response to demands for transparent decision support. Arrieta et al. [17] provided a comprehensive taxonomy of XAI concepts, methods, and challenges, establishing that explainability is a fundamental prerequisite for accountable and trustworthy deployment in high-stakes domains. Lundberg and Lee [18] introduced SHAP as a unified framework for feature attribution. Across these contributions, explainability is predominantly operationalized at the feature attribution level rather than at the level of cross-representational reasoning consistency. The present paper contributes a topology-aware validity condition for the structural reasoning layer itself, complementing existing explainability frameworks by establishing when structural explanations are epistemically reliable.

2.6 Temporal and Production-Traffic Validity in Intrusion Detection

Recent intrusion detection research increasingly recognizes that network behavior is temporally structured rather than adequately represented by isolated flow records. Large-scale operational datasets and dynamic graph anomaly detection studies show that temporal variability, entity persistence, and evolving graph states affect anomaly detection validity and deployment realism.

Koumar et al. [16] introduced CESNET-TimeSeries24, derived from 40 weeks of traffic involving approximately 275,000 active IP addresses in the CESNET3 ISP network, explicitly motivating evaluation under realistic temporal and entity-diverse conditions — properties largely absent from small laboratory testbeds where source populations are constrained and topology remains comparatively stable.

Dynamic graph anomaly detection research has emphasized the importance of modeling evolving relationships. Ares-Robledo et al. [11] reviewed temporal graph neural network approaches for anomaly detection, identifying cross-domain validation and deployment readiness as continuing challenges.

However, this literature primarily focuses on learning temporal graph representations. It does not examine the prior validity question of whether structural indicators remain semantically stable across observation windows and production traffic regimes — the question addressed in § 5.9.

Production-web validation also introduces label uncertainty absent from curated benchmarks. Reputation-based labeling is operationally useful but must be interpreted cautiously, as reputation scores are inferred indicators shaped by reporting practices, deployment coverage, and context-dependent interpretations of automated or gray-zone activity [19]. For this reason, the production validation treats AbuseIPDB-derived labels as external reputation indicators rather than definitive ground truth.

2.7 Research Gap

Existing graph-based intrusion detection research primarily evaluates how effectively a detection model can exploit structural information once a communication graph has been constructed. This model-centric paradigm leaves a prior validity question unresolved: whether the underlying topology itself provides structural differentiation that is security-selective rather than merely statistically heterogeneous. Structural prominence may arise from malicious scanning, coordinated attack behavior, or command-and-control activity, but it may also arise from legitimate gateways, servers, crawlers, or other high-connectivity infrastructure. Consequently, structural inequality alone does not establish that graph-derived prominence constitutes a meaningful intrusion signal.

This distinction exposes a gap not explicitly addressed in existing graph-based, hybrid, and centrality-based IDS research. Prior work does not separate two properties that may vary independently across deployment environments: **structural differentiation**, referring to whether the topology contains distinguishable connectivity regimes, and **prominence isolation**, referring to whether the structurally prominent region is sufficiently separated from benign high-connectivity entities to support attack-selective reasoning. Without this distinction, an environment may appear structurally favorable according to aggregate topology statistics while providing little or no operational discrimination.

A related validity problem arises when structural differentiation exists but conventional percentile-based thresholds fail to reveal it. Large peripheral entity populations can cause structural thresholds to collapse even when a small, strongly isolated structural tail remains present. Threshold failure must therefore be distinguished from genuine absence of structural signal. In addition, production environments introduce a temporal dimension: entity accumulation, persistence patterns, and event-driven traffic shifts may alter the relationship between structural differentiation and prominence isolation over time.

The unresolved problem is therefore not simply whether graph-based intrusion detection can achieve high detection performance, but **under what topological conditions structural prominence constitutes a valid and attack-selective signal for intrusion reasoning in the first place**. This study addresses that problem by separating Graph Structural Differentiability (GSD) from Prominence Isolation (PI), empirically

testing their dissociation across heterogeneous environments, characterizing topology-dependent threshold-collapse regimes, and operationalizing these findings through a pre-deployment validity assessment.

Accordingly, the study is guided by the following central research question:

RQ: Under what topological conditions does structural prominence constitute a valid, attack-selective signal for graph-based intrusion detection rather than merely reflecting benign structural heterogeneity?

This question is examined through two subsidiary questions:

RQ1

Can structural differentiation and attack-selective prominence dissociate across heterogeneous network environments?

RQ2

When conventional structural thresholds fail, can label-free topology diagnostics distinguish between genuine absence of structural signal and recoverable threshold collapse?

3. Methodology

3.1 Framework Overview

The proposed framework evaluates graph-based structural validity before interpreting downstream detection performance. It separates two complementary dimensions that are often implicitly conflated: **structural differentiation** and **prominence isolation**. Structural differentiation concerns whether the topology contains connectivity regimes that are distinguishable from the background population; prominence isolation concerns whether the structurally prominent region is sufficiently isolated to plausibly support attack-selective reasoning rather than being dominated by routine high-connectivity infrastructure.

These dimensions are operationalized through Graph Structural Differentiability (GSD) and Prominence Isolation (PI). GSD characterizes the global and tail-level structure of the source out-degree distribution using multiple complementary indicators. PI evaluates the isolation characteristics of the structurally prominent region using label-free diagnostics during pre-deployment assessment, with ground-truth information introduced only for post-hoc validation where available. The framework therefore does not assume that high structural inequality implies useful intrusion discrimination; rather, it explicitly tests whether structural differentiation and prominence isolation align within the target environment.

The methodology proceeds in two phases. **Phase 1** evaluates the topology prior to detection by computing GSD indicators, identifying any Threshold Collapse regime, and assessing PI under the

applicable structural threshold. Threshold Collapse is treated as a diagnostic mechanism for distinguishing genuine structural degeneracy from recoverable threshold failure. The outcome of Phase 1 determines whether structural reasoning is considered supported, recoverable through topology-adaptive thresholding, or unsuitable for operational interpretation.

Phase 2 evaluates the structural channel together with behaviorally independent detection paradigms and a convergence gate. This phase does not define structural validity; instead, it provides empirical evidence for whether the structural regimes identified in Phase 1 correspond to attack-selective detection utility and cross-paradigm consistency. Ground-truth labels are withheld from graph construction, structural thresholding, and label-free validity diagnostics, and are introduced only where required for supervised reference modeling and post-hoc evaluation.

Let $\mathbf{G} = (\mathbf{V}, \mathbf{E})$ denote a directed communication graph constructed from network flow records, where $\mathbf{V}$ is the set of source entities and $\mathbf{E} \subseteq \mathbf{V} \times \mathbf{V}$ is the set of directed edges representing unique destination connections. The source out-degree of entity $v \in \mathbf{V}$ is:

$$d_{\text{out}}(v) = |\{u \in \mathbf{V} : (v, u) \in \mathbf{E}\}|$$

All structural validity analyses are derived from the empirical distribution of d_{out} across the source-entity population and are computable without attack labels except where explicitly identified as post-hoc validation.

3.2 Graph Structural Differentiability (GSD)

Definition. Graph Structural Differentiability (GSD) is a multi-indicator characterization of structural heterogeneity in the source out-degree distribution. It describes whether the topology contains distinguishable connectivity regimes, but it does not by itself determine whether structurally prominent entities constitute a security-relevant signal. High GSD therefore indicates the presence of structural differentiation that may support graph-based reasoning, whereas operational validity requires complementary assessment of Prominence Isolation (PI).

GSD is operationalized through four complementary indicators, each capturing a distinct aspect of distributional heterogeneity:

Gini Coefficient (G_{ini}). The Gini coefficient of the out-degree distribution:

$$G_{\text{ini}} = (2 \cdot \sum i \cdot \tilde{d}_{\text{out}}(i) - (n+1) \cdot \sum \tilde{d}_{\text{out}}(i)) / (n \cdot \sum \tilde{d}_{\text{out}}(i))$$

where $\tilde{d}_{\text{out}}(1) \leq \dots \leq \tilde{d}_{\text{out}}(n)$ is the sorted out-degree sequence and $n = |\mathbf{V}|$. Values approaching 1.0 indicate high structural inequality; values near 0.0 indicate near-uniform connectivity.

Coefficient of Variation (CV). $CV = \sigma(d_{\text{out}}) / \mu(d_{\text{out}})$ measures distributional spread relative to the mean, capturing the degree to which tail entities dominate the distribution.

P95/Median Ratio (R_{95}). The ratio of the 95th percentile out-degree to the median:

$$R_{95} = P_{95}(d_{out}) / \text{median}(d_{out})$$

Large values of R_{95} indicate that entities in the structural tail exhibit connectivity patterns qualitatively distinct from the majority population. Values near 1.0 indicate distributional homogeneity.

Threshold Collapse Indicator (TC). Threshold Collapse is an operational condition in which percentile-based structural thresholds fail to produce meaningful reduction in the structural candidate set across increasing percentile levels. TC is operationally indicated when the structural candidate sets generated at P90, P95, and P97 exhibit near-identical cardinality. Formally:

$$TC = 1 \text{ if } |P_{97}(d_{out}) - P_{90}(d_{out})| / \max(P_{97}(d_{out}), 1) < \delta_{TC}$$

where $\delta_{TC} = 0.05$ is the collapse tolerance. TC is complementarily assessed through the Maximum Gap Ratio (MGR, defined in §3.3): low MGR in the structural tail provides convergent evidence of genuine distributional collapse.

GSD interpretation. GSD indicators are interpreted jointly rather than through fixed universal thresholds. Their role is to characterize the presence and form of structural differentiation, not to establish attack selectivity. The empirical results of this study show that high aggregate structural inequality can coexist with low Prominence Isolation, while apparent threshold collapse can coexist with a recoverable isolated tail. GSD therefore represents the structural-differentiation dimension of the validity assessment and must be interpreted together with PI before conclusions about structural intrusion utility are drawn.

3.3 Prominence Isolation (PI)

Definition. Prominence Isolation (PI) characterizes whether the structurally prominent tail is sufficiently isolated from the surrounding entity population to plausibly support attack-selective structural reasoning. PI is not a direct label-free measure of maliciousness; rather, it evaluates topological conditions associated with selective prominence before ground-truth labels are introduced. High PI indicates a sparse and structurally separated prominent region, whereas low PI indicates a dense or weakly separated tail in which benign high-connectivity infrastructure may dominate structural prominence. Attack selectivity is therefore inferred operationally from label-free PI diagnostics and evaluated empirically through post-hoc ground-truth validation where labels are available.

PI is operationalized through two label-free operational diagnostics and one post-hoc validation metric:

Tail Density (TD) – label-free. The proportion of entities occupying the structural tail:

$$TD = |\{v \in V : d_{out}(v) \geq \tau\}| / |V|$$

where τ is the applicable structural threshold. TD is interpreted relative to the thresholding regime identified during TC assessment. Low TD indicates a sparse, well-isolated tail; high TD indicates a

densely populated tail in which structural prominence provides limited discriminative utility.

Maximum Gap Ratio (MGR) — label-free. The largest consecutive ratio in the sorted descending out-degree sequence of the top- k structurally prominent entities, computed after excluding peripheral entities with $d_{\text{out}} \leq 1$:

$$MGR = \max_{\{i=1, \dots, k-1\}} \tilde{d}_{\text{out}}(i) / \tilde{d}_{\text{out}}(i+1)$$

where $\tilde{d}_{\text{out}}(1) \geq \dots \geq \tilde{d}_{\text{out}}(k)$ is sorted in descending order over entities with $d_{\text{out}} > 1$, and k is the analytical window ($k = 30$ in this study). Excluding singleton-degree peripherals prevents MGR from being dominated by trivial floor effects at the boundary between sparse tail entities and the low-degree bulk population. High MGR ($> 5.0\times$) indicates a clear natural discontinuity separating the prominent region from the remaining population; low MGR ($< 2.0\times$) indicates a continuous, undifferentiated tail with no recoverable structural break. These ranges are empirical operational heuristics derived from the evaluated environments rather than theoretically universal constants.

Tail Purity (TP) — post-hoc, requires ground-truth. The proportion of attacker entities within the structural tail:

$$TP = |\{v : d_{\text{out}}(v) \geq \tau, v \in \text{Attacker}\}| / |\{v : d_{\text{out}}(v) \geq \tau\}|$$

TP is used exclusively for post-hoc validation when ground-truth labels are available and does not form part of the pre-deployment diagnostic protocol.

PI interpretation. PI complements GSD by addressing a different validity dimension. GSD determines whether structural differentiation exists, whereas PI assesses whether the prominent region is sufficiently isolated for that differentiation to plausibly support selective intrusion reasoning. PI therefore does not assert maliciousness from topology alone. Its pre-deployment role is diagnostic: low TD and a pronounced tail discontinuity provide label-free evidence of structural isolation, while Tail Purity is used only after labels become available to test whether that isolation corresponds to attacker enrichment. The empirical analysis in §5 evaluates this distinction directly.

3.4 Threshold Collapse: Static Forms in Benchmark Environments

Full-scale analysis across the evaluated environments shows that failure of percentile-based structural thresholds does not have a single interpretation. Threshold Collapse may reflect either a genuine absence of recoverable structural differentiation or a thresholding artifact produced by the composition of the entity population. Distinguishing these cases is necessary because apparent structural failure should not be interpreted as evidence of topology-level invalidity unless the underlying tail structure is also examined.

Homogeneous Collapse occurs when the out-degree distribution is genuinely near-uniform: the vast majority of entities exhibit similar connectivity, structural outliers are absent or negligible, and MGR is

near 1.0× throughout the top-k sequence. In homogeneous collapse environments, no threshold mechanism can produce meaningful structural separation because the structural signal is absent from the topology. UNSW-NB15 and NF-UNSW-NB15-v2 are characteristic examples: all entities exhibit out-degrees within a narrow range (max out-degree $\approx 10\text{--}11$), and MGR $\approx 1.0\times$ confirms genuine distributional homogeneity.

Population-Expansion Collapse occurs when the entity population is dominated by peripheral low-degree nodes — characteristic of large IoT environments and enterprise networks with many minimally-connected endpoints — causing percentile thresholds to degenerate to near-zero values despite the continued presence of extreme structural outliers in the tail. NF-ToN-IoT-v2 is a characteristic example: with 23,079 entities of which more than 99% exhibit out-degree = 1, the P95 threshold collapses to 1, yielding $S = V$. Yet the top entities exhibit out-degrees exceeding 5,000, and MGR = 12.47× (computed over entities with $d_{\text{out}} > 1$) confirms the presence of a strong natural discontinuity separating a small cluster from the peripheral population.

These two forms are distinguished through the joint evaluation of TC and MGR:

Table 1. Diagnostic Interpretation of Static Threshold Collapse Forms

TC	MGR	Form	Operational Implication
0	Any	No collapse	Percentile threshold operationally stable; proceed with PI assessment
1	$< 2.0\times$	Homogeneous	Structural reasoning unsuitable; rely on behavioral channels
1	$\geq 5.0\times$	Population-expansion	Gap-based threshold applicable; structural signal recoverable

Note: TC = 0 indicates threshold stability only and does not imply structural channel validity. PI assessment remains required regardless of TC value, as environments with stable percentile thresholds may nonetheless exhibit insufficient Prominence Isolation.

A third, temporally induced form of Threshold Collapse is identified in the production validation analysis (§5.9.5), where a legitimate event-driven traffic surge temporarily degrades structural discriminability despite subsequent full recovery.

Interpretation. Threshold Collapse is treated as a diagnostic mechanism rather than as an independent criterion of structural validity. Homogeneous collapse indicates that the topology provides insufficient structural differentiation for meaningful structural filtering. Population-expansion collapse, by contrast, indicates that structural differentiation remains present but is obscured by percentile degeneration caused by a large peripheral population. In the latter case, structural utility remains potentially recoverable and must be evaluated through topology-adaptive thresholding followed by PI assessment.

Accordingly, TC identifies the form of threshold failure, while GSD and PI determine the broader interpretation of structural validity.

3.5 Topology-Adaptive Gap-Based Thresholding

When Threshold Collapse is classified as population-expansion rather than homogeneous, the structural signal is treated as potentially recoverable rather than absent. In this regime, conventional percentile thresholds are replaced by a topology-adaptive gap-based threshold derived from a natural discontinuity in the upper out-degree distribution. The purpose of this procedure is not to redefine structural validity, but to recover a candidate structural region that can subsequently be evaluated through PI.

Threshold Derivation. Two quantities are distinguished: (i) MGR, the maximum gap ratio, which serves as a PI diagnostic confirming that a structural break exists in the tail; and (ii) τ_{gap} , the operational threshold derived from a meaningful gap position that produces a non-trivial structural set.

Given the sorted descending out-degree sequence over entities with $d_{\text{out}} > 1$, the set of structurally significant gaps is:

$$\Omega = \{i : \tilde{d}_{\text{out}}(i) / \tilde{d}_{\text{out}}(i+1) > \gamma_{\text{min}}\}$$

where $\gamma_{\text{min}} = 3.0\times$ is the minimum gap ratio for operational significance. The gap-based threshold τ_{gap} is set at the upper bound of the gap in Ω that produces the smallest non-trivial structural set:

$$\tau_{\text{gap}} = \tilde{d}_{\text{out}}(i^*) \text{ where } i^* = \operatorname{argmin}_{\{i \in \Omega\}} |S_{\geq i}| \text{ subject to } |S_{\geq i}| \geq k_{\text{min}}$$

where $S_{\geq i} = \{v : d_{\text{out}}(v) \geq \tilde{d}_{\text{out}}(i)\}$ and k_{min} is the minimum structural set size ($k_{\text{min}} = 2$ in this study). τ_{gap} corresponds to the out-degree value at the upper bound of the selected gap — that is, the value immediately before the structural discontinuity, ensuring that the structural set S contains entities above the natural break rather than below it.

In NF-ToN-IoT-v2, MGR = $12.47\times$ at position $1 \rightarrow 2$ would yield $|S| = 1$ entity, violating the k_{min} constraint. The next significant gap occurs at position $11 \rightarrow 12$ (out-degree ratio $214 \rightarrow 49$, gap ratio $4.37\times$), yielding $\tau_{\text{gap}} = 214$ and $|S| = 11$ entities. This corrected threshold produces Hybrid Precision = 0.909 and Lift $\approx 677\times$, where the conventional P95 threshold produces Precision = 0.001 and Lift $\approx 1.0\times$.

The structural set under gap-based thresholding is defined as:

$$S = \{v \in V : d_{\text{out}}(v) \geq \tau_{\text{gap}}\}$$

Interpretation. Gap-based thresholding is therefore a recovery mechanism for a specific threshold-failure regime, not evidence of structural validity by itself. A recovered structural set remains operationally meaningful only if subsequent PI assessment indicates sufficient prominence isolation. This distinction prevents successful threshold separation from being conflated with attack-selective structural utility.

3.6 Structural and Behavioral Detection Channels

Structural Channel. The structural channel flags entities whose source out-degree exceeds the applicable threshold — either the conventional P95 threshold in non-collapse environments, or τ_{gap} in population-expansion collapse environments:

$$S(a) = 1 \text{ if } d_{\text{out}}(a) \geq \tau, \text{ else } 0$$

The structural channel is computationally independent of flow-level behavioral features, deriving its signal exclusively from graph-topological properties.

Behavioral Channel. The behavioral channel operates over flow-level statistical features — including flow duration, byte counts, packet counts, and throughput measures. A source entity is flagged behaviorally if at least one observed flow exceeds the P95 threshold of the corresponding feature distribution computed over benign-labeled flows:

$$B(a) = 1 \text{ if } \exists f \in \text{flows}(a) : f_j \geq P_{95}(F_j^{\text{benign}}) \text{ for some feature } j$$

The behavioral channel is operationalized across three methodologically distinct paradigms to assess cross-paradigm structural invariance: threshold-based detection (B), Isolation Forest (IF), and Random Forest (RF). Features are normalized using RobustScaler prior to model training to mitigate the influence of extreme values characteristic of high-volume network traffic.

Hybrid Convergence Gate. The hybrid gate requires corroboration from both channels:

$$\text{Alert}(a) = S(a) \wedge B(a)$$

Cross-Paradigm Validation. Cross-paradigm analysis is used as a robustness test of the topology-validity interpretation rather than as an independent validity criterion. The objective is to examine whether structurally prominent entities identified under a given GSD-PI regime remain aligned with behaviorally anomalous or malicious entities across methodologically distinct detection paradigms. Consistent structural inclusion across behavioral methods is therefore interpreted as supporting evidence only when the topology has already exhibited sufficient structural differentiation and prominence isolation.

Conversely, subset agreement is not considered meaningful by itself. In low-PI or structurally degenerate environments, a structural set may be embedded within broad behavioral anomaly sets simply because those sets contain many benign entities. Such agreement represents convergent non-discrimination rather than attack-selective corroboration. Cross-paradigm results are therefore interpreted conditionally on the structural-validity regime established through GSD, Threshold Collapse assessment, and PI.

Three methodologically distinct behavioral paradigms were employed, representing a spectrum from profile-based statistical thresholding to fully supervised classification.

(i) Threshold-based detection (B): entities are flagged if any observed flow exceeds the P95 threshold of the benign flow distribution for at least one behavioral feature. Thresholds are derived from flows identified as benign; attack labels are not used during threshold computation or entity flagging.

(ii) Isolation Forest (IF): a fully unsupervised anomaly detector trained on the complete flow-level feature matrix without labels, with contamination parameters set to 0.05 and 0.03 for P95 and P97 evaluations respectively.

(iii) Random Forest (RF): a supervised classifier trained and evaluated within each environment using ground-truth attack labels from the corresponding environment-specific dataset. RF operates as an in-sample supervised reference paradigm rather than a cross-dataset transfer model; its role is to assess whether structurally prominent entities align with entities classified as malicious by a fully supervised classifier within the same environment, rather than to demonstrate unsupervised generalization.

Attack labels are withheld throughout structural channel construction and convergence gate operation. Behavioral thresholds for the B channel are derived from benign-labeled flows; attack-specific labels are introduced exclusively for RF training and post-hoc validation of detection outcomes.

Cross-paradigm evaluation therefore spans a deliberate methodological spectrum from statistical thresholding to unsupervised anomaly detection and supervised classification. The resulting subset relationships are used to test whether the structurally selected entities remain aligned with independently derived behavioral evidence across these paradigms. These relationships are not treated as validity conditions themselves; their epistemic meaning depends on the GSD–PI regime in which they occur.

In environments with sufficient GSD and PI, the subset relationship $S \subset B \cap RF \cap IF$ was empirically observed to hold. This was confirmed in WTM2021, which provided the primary validation in the companion study, and in NF-ToN-IoT-v2, validated at full scale across 16.9 million flow records, establishing that the observed relationship is not an artifact of sampling procedures.

In environments with insufficient PI, the structural set converges toward benign infrastructure entities. In NF-CSE-CIC-IDS2018-v2, the gap-based structural set contained a single benign entity with out-degree 2,525 – the highest-degree entity in a network of 224,568 source entities across 18.9 million flow records – yielding $S \subset B$ trivially while $S \subset RF$ failed, reflecting structural misalignment rather than invariance.

In homogeneous collapse environments (UNSW-NB15 and NF-UNSW-NB15-v2), the structural channel loses discriminative utility entirely. Cross-paradigm invariance in these environments becomes epistemically meaningless: agreement between S and behavioral channels reflects convergent non-discrimination rather than reliable attack-selective corroboration. Full empirical results across all five environments are reported in §5.

3.7 Pre-Deployment Diagnostic Protocol

The pre-deployment diagnostic protocol operationalizes GSD and PI assessment as a sequential decision procedure executable from unlabeled network traffic prior to deployment.

1. **Step 1 — Observation Window Specification.** In production deployments, specify the observation window before graph construction. Short windows (1–3 days) preserve higher Tail Purity and are appropriate when broad structural coverage is the operational objective; longer windows (7+ days) yield higher Precision@Top-k and are appropriate when prioritizing structurally persistent entities for investigation. Window length should be aligned with operational objectives (§5.9).
2. **Step 2 — Graph Construction.** Construct $G = (V, E)$ from available network traffic. Compute $d_{\text{out}}(v)$ for all $v \in V$.
3. **Step 3 — GSD Assessment.** Compute Gini, CV, R_{95} , and TC jointly. Indicators are evaluated in combination; no fixed universal thresholds apply.
4. **Step 4 — TC Form Identification.** If $TC = 1$, compute MGR over top-30 entities with $d_{\text{out}} > 1$. $MGR \geq 5.0\times$: population-expansion collapse → apply gap-based threshold (§3.5). $MGR < 2.0\times$: homogeneous collapse → flag environment as structurally unsuitable; rely on behavioral channels alone.
5. **Step 5 — Threshold Selection.**
 - $TC = 0$ → apply the conventional P95 threshold.
 - Population-expansion collapse → apply τ_{gap} .
 - Homogeneous collapse → structural reasoning unsuitable; rely on behavioral channels alone.
6. **Step 6 — PI Assessment.** Compute TD and MGR under the applicable structural threshold. Low TD together with a pronounced structural gap provides label-free evidence of prominence isolation and supports continued structural evaluation. High TD or weak tail separation indicates an increased risk that structural prominence is dominated by benign high-connectivity infrastructure. Where labels are available for retrospective evaluation, Tail Purity is used only as a post-hoc test of whether the isolated structural region is attacker-enriched.
7. **Step 7 — Cross-Channel Interpretation.** Interpret structural–behavioral agreement only after the topology has been characterized through GSD, Threshold Collapse assessment, and PI. High agreement in low-PI or structurally degenerate environments may reflect convergent non-discrimination rather than reliable attack-selective corroboration. RCI and subset relationships are therefore treated as supporting evidence whose meaning is conditional on the previously established topology-validity regime.

4. Experimental Setup

4.1 Evaluation Environments

Experiments were conducted across five heterogeneous network environments spanning enterprise, IoT, and laboratory network topologies, selected to represent qualitatively distinct topological regimes

enabling systematic GSD and PI evaluation.

WTMC2021 is a derivative of CIC-IDS2017, retaining the entity-level communication structure suitable for structural channel evaluation. It contains 19,060 unique source entities with a strongly right-skewed out-degree distribution (Gini = 0.799; P95/Median = 47,144×), characterized by a small number of structurally isolated high-degree attacker entities against a low-degree bulk population. It serves as the primary high-GSD/high-PI reference environment throughout the study.

NF-ToN-IoT-v2 is derived from the ToN-IoT benchmark, a multi-protocol IoT dataset incorporating diverse attack categories. Full-scale analysis covers 16,940,496 flow records yielding 23,079 source entities. Over 99% of entities exhibit out-degree = 1, producing population-expansion threshold collapse; however, a small cluster of high-degree entities achieves extreme structural prominence (MGR = 12.47×), confirming a recoverable structural signal under gap-based thresholding ($\tau = 214$; structural set of 11 entities).

NF-CSE-CIC-IDS2018-v2 is derived from CIC-IDS2018, incorporating DDoS, DoS, brute-force, bot, and infiltration traffic. Full-scale analysis covers 18,893,708 flow records yielding 224,568 entities. Despite high structural inequality (Gini = 0.812), all six highest-degree entities are benign infrastructure nodes, with the first attacker appearing at structural rank 7. This environment serves as the primary low-PI reference case.

NF-UNSW-NB15-v2 is the NF-UQ-NIDS-v2 representation of UNSW-NB15, generated under IXIA PerfectStorm constraints. With only 40 unique source entities and a near-uniform out-degree distribution (max = 10), it exhibits homogeneous threshold collapse and serves as the primary homogeneous collapse reference.

UNSW-NB15 (original format) contains 47 source entities under the same IXIA PerfectStorm conditions, confirming that record-volume expansion does not restore structural separation in synthetically generated traffic. It establishes the lower bound of structural discriminability in the evaluation suite.

Table 2 summarizes topological characteristics. The production web validation environment analyzed in § 5.9 is reported separately, as its purpose is temporal validation rather than inclusion in the static topology taxonomy.

Table 2
Evaluation Environments and Their Experimental Roles

Environment	Flow Records	Entities	Attackers	Base Rate	Role in Study
WTMC2021	—	19,060	8	0.042%	High-GSD / high-PI reference
NF-ToN-IoT-v2	16,940,496	23,079	31	0.134%	Population-expansion collapse / recoverable PI
NF-CSE-CIC-IDS2018-v2	18,893,708	224,568	648	0.289%	Low-PI infrastructure-dominated reference
NF-UNSW-NB15-v2	—	40	—	—	Homogeneous collapse reference
UNSW-NB15	—	47	—	—	Homogeneous low-discriminability baseline

4.2 Dataset Construction and Graph Construction

Graph construction. For each environment, a directed communication graph $G = (V, E)$ was constructed from the available flow records by treating unique source IP addresses as source entities and computing source out-degree as the number of unique destination IP addresses contacted by each source. Edge construction was based on unique (source IP, destination IP) pairs rather than individual flow records, ensuring that out-degree reflects the breadth of an entity's communication footprint rather than its traffic volume. All graph construction was performed without reference to ground-truth labels.

Full-scale vs. sampled evaluation. GSD indicators for WTMC2021 were computed from the full available dataset; structural detection performance results for this environment are attributed to the companion study. NF-ToN-IoT-v2 and NF-CSE-CIC-IDS2018-v2 structural analyses were conducted on complete datasets of 16,940,496 and 18,893,708 flow records respectively, using Neo4j Desktop without node or edge count restrictions. UNSW-NB15 and NF-UNSW-NB15-v2 structural analyses were conducted on full available datasets.

Sampling effects on GSD. An important methodological observation emerged from comparing sampled and full-scale entity distributions for NF-ToN-IoT-v2. The stratified sampling procedure used in preliminary analysis produced a 108-entity sample exhibiting Gini = 0.845 and no threshold collapse. Full-scale analysis of 23,079 entities revealed Gini = 0.271 and population-expansion threshold collapse. This divergence reflects the sampling procedure's inherent tendency to over-represent entities with diverse traffic patterns, which correlates with attack-associated behavior. This finding motivates the recommendation that GSD assessment be conducted on full-scale entity populations wherever computationally feasible, as sampled GSD values may not reliably characterize the topological regime of the full deployment environment.

Behavioral feature construction. Behavioral channel evaluation utilized seven flow-level statistical features available across all evaluation environments: flow duration, inbound and outbound byte counts, inbound and outbound packet counts, and source-to-destination and destination-to-source second-byte rates. Features were normalized using RobustScaler prior to threshold computation and model training to mitigate the influence of extreme values characteristic of high-volume network flows.

4.3 Evaluation Protocol

Label-withheld detection protocol. Graph construction, structural thresholding, GSD assessment, PI diagnostics, and convergence-gate operation are performed without access to ground-truth attack labels. Labels are introduced only where explicitly required for supervised Random Forest reference modeling and for post-hoc evaluation of structural and hybrid detection outcomes. This design preserves the pre-deployment character of the validity framework: topology suitability is assessed without using attack labels to construct or calibrate the structural signal.

Cross-paradigm evaluation. The behavioral channel was evaluated under three methodologically distinct paradigms representing a spectrum from profile-based statistical thresholding to fully supervised classification, as described in § 3.6. Briefly: threshold-based detection (B) flags entities based on benign flow baselines without attack labels; Isolation Forest (IF) operates as a fully unsupervised anomaly detector; and Random Forest (RF) serves as an in-sample supervised reference, trained and evaluated independently within each environment to avoid cross-environment distribution leakage. Cross-paradigm invariance was assessed by verifying the subset relationship $S \subset B \cap IF \cap RF$ at the entity level across threshold levels P95 and P97.

Performance metrics. Detection performance is reported through four primary metrics. Precision and Recall are computed at the entity level against post-hoc ground-truth labels. Lift is defined as the ratio of observed precision to the unconditional base rate of attackers in the evaluation universe, providing a base-rate-normalized measure of structural enrichment that supports comparison across environments with substantially different attacker prevalence rates. The Reasoning Consistency Index (RCI) is defined as the Jaccard similarity between the structural set S and the behavioral anomaly set B , measuring the degree of cross-channel agreement. As established in § 3.7, RCI is interpreted as epistemically meaningful only in environments with confirmed GSD and PI sufficiency; high RCI in collapse or low-PI environments reflects convergent non-discrimination rather than reliable attack-selective corroboration.

Full-scale cross-paradigm validation. For NF-ToN-IoT-v2, cross-paradigm invariance validation was conducted on the complete 16,940,496-row dataset with gap-based structural thresholding, providing a full-scale replication of the cross-paradigm invariance finding. For NF-CSE-CIC-IDS2018-v2, full-scale structural and cross-paradigm analysis was conducted on the complete 18,893,708-row dataset, providing empirical confirmation that the low-PI characterization established through entity-level analysis holds under full behavioral channel evaluation.

5. Results

5.1 GSD and PI Characterization Across Environments

The cross-environment analysis reveals that structural differentiation and prominence isolation do not vary as a single property. Environments with comparable or even greater aggregate structural inequality can exhibit sharply different levels of attack-selective structural utility, while apparent percentile collapse can coexist with a recoverable isolated tail. Table 3 therefore reports GSD and PI indicators jointly, allowing the evaluated environments to be distinguished according to both the presence of structural differentiation and the isolation characteristics of the structurally prominent region.

Table 3
GSD and PI Diagnostic Indicators Across Evaluation Environments

Environment	Gini	P95/Med	TC	MGR	TD (P95)	TP (P95)	GSD	PI
WTMC2021	0.799	47,144×	No	> 500×	0.016%	100%	High	Very High
NF-ToN-IoT-v2	0.271	1.0×	Yes (Pop-Exp)	12.47×	0.048%*	90.9%*	Mixed (Pop-Exp)	High*
NF-CSE-CIC-IDS2018-v2	0.812	22×	No	1.19×	5.15%	2.6%	Medium	Low
NF-UNSW-NB15-v2	0.359	1.0×	Yes (Hom)	1.0×	53.8%	—	Low	Low
UNSW-NB15	0.256	1.18×	Yes (Hom)	1.1×	40.4%	—	Low	Low

*Under gap-based threshold ($\tau_{\text{gap}} = 214$); conventional P95 yields TD = 100% (full population collapse).

The comparatively low Gini value observed in NF-ToN-IoT-v2 reflects the dominance of the entity population by peripheral degree-1 nodes rather than the absence of structurally isolated outliers. In this regime, aggregate inequality metrics underestimate tail isolation, motivating the complementary PI diagnostics that reveal an MGR of 12.47×

and a Tail Purity of 90.9% under gap-based thresholding. Three topological regimes are empirically distinguished. WTMC2021 exhibits the reference high-GSD, high-PI configuration: extreme distributional inequality with structurally isolated high-degree entities confirmed post-hoc as attackers. NF-ToN-IoT-v2 exhibits a structurally distinct regime — population-expansion collapse with recoverable structural signal — where conventional GSD indicators are depressed by the large peripheral population yet extreme structural outliers with MGR = 12.47×

collapse environments exhibit homogeneous collapse with $MGR \approx 1.0\times$ across the top-k sequence, confirming the absence of recoverable structural signal.

5.2 Structural Channel Performance Under Standard Thresholding

Table 4 presents structural channel (S-only) performance under conventional P95 and P97 thresholds across all environments.

Table 4
Structural Channel Performance — Conventional Percentile Thresholds

Environment	Threshold	S Size	TP	FP	Precision	Recall	Lift
WTMC2021	P95	3	3	0	1.000	0.375	2,383×
NF-ToN-IoT-v2	P95	23,079	31	23,048	0.001	1.000	1.0×
NF-ToN-IoT-v2	P97	23,079	31	23,048	0.001	1.000	1.0×
NF-CSE-CIC-IDS2018-v2	P95	11,571	301	11,270	0.026	0.465	9.0×
NF-CSE-CIC-IDS2018-v2	P97	6,829	257	6,572	0.038	0.397	13.0×
NF-UNSW-NB15-v2	P95	14	—	—	0.286	—	—
UNSW-NB15	P95	19	—	—	0.211	—	—

Under standard percentile thresholding, performance diverges substantially across topological regimes. WTMC2021 achieves Precision = 1.000 with zero false positives, consistent with its high-GSD, high-PI characterization. NF-ToN-IoT-v2 exhibits complete structural collapse under both P95 and P97: the threshold degenerates to 1, flagging all 23,079 entities and producing Lift $\approx 1.0\times$, equivalent to random selection. NF-CSE-CIC-IDS2018-v2 achieves moderate lift (9.0–13.0×) but low absolute precision (0.026–0.038), reflecting the dilution of the structural tail by benign high-connectivity infrastructure. The collapse environments exhibit limited structural utility regardless of threshold level.

5.3 Population-Expansion Collapse Recovery: NF-ToN-IoT-v2

Table 5 compares structural channel performance under conventional and gap-based thresholding for NF-ToN-IoT-v2, demonstrating that population-expansion collapse is a threshold artifact rather than a genuine absence of structural signal.

Table 5
NF-ToN-IoT-v2: Conventional vs. Gap-Based Threshold Comparison

Threshold	Criterion	S Size	TP	FP	Precision	Recall	Lift
P95 (conventional)	$d_{out} \geq 1$	23,079	31	23,048	0.001	1.000	1.0×
P97 (conventional)	$d_{out} \geq 1$	23,079	31	23,048	0.001	1.000	1.0×
Gap-based (τ_{gap})	$d_{out} \geq 214$	11	10	1	0.909	0.323	677×

The gap-based threshold recovers a structural set of 11 entities, of which 10 are confirmed attackers post-hoc (Precision = 0.909, Lift = 677×). The single false positive is entity 192.168.1.180 (out-degree = 225), whose out-degree falls just above the gap threshold and was later identified as a benign high-connectivity device. The structural set is stable across P95 and P97 behavioral threshold levels, confirming threshold-invariant structural isolation.

The top entity, 192.168.1.190, exhibits out-degree = 5,161 — a separation of nearly three orders of magnitude from the bulk population (out-degree = 1) and producing the maximum observed consecutive gap ratio (12.47×) within the structural tail. This extreme structural prominence is characteristic of a scanning or command-and-control source operating across a large fraction of the observed IoT device population.

5.4 Low-PI Structural Failure: NF-CSE-CIC-IDS2018-v2

Table 6 presents structural channel performance across multiple threshold strategies for NF-CSE-CIC-IDS2018-v2, demonstrating that PI insufficiency produces structural filtering failure independent of threshold selection.

Table 6
NF-CSE-CIC-IDS2018-v2: Structural Channel Across Threshold Strategies

Threshold	Criterion	S Size	TP	FP	Precision	Recall	Lift
P95 (percentile)	$d_{out} \geq 22$	11,571	301	11,270	0.026	0.465	9.0×
P97 (percentile)	$d_{out} \geq 31$	6,829	257	6,572	0.038	0.397	13.0×
Gap-based (τ_{gap})	$d_{out} \geq 2,525$	1	0	1	0.000	0.000	0×
Absolute ≥ 500	$d_{out} \geq 500$	392	16	376	0.041	0.025	14.1×
Top-11 (matched)	Top 11 entities	11	1	10	0.091	0.002	31.5×

The gap-based threshold selects a single entity — 172.31.67.62 with out-degree = 2,525, confirmed post-hoc as a benign internal infrastructure node. This entity is the highest-degree node in a network of 224,568 source entities across 18.9 million flow records, yet exhibits no malicious behavior, illustrating the PI failure mode directly: the structurally dominant entity is a legitimate high-connectivity gateway rather than an anomalous actor. No threshold strategy produces precision exceeding 0.091 in this

environment, establishing that the structural failure is not a consequence of threshold selection but of the topology's fundamental PI insufficiency.

5.5 Cross-Paradigm Structural Invariance

Table 7 presents cross-paradigm invariance results across all five environments at P95 and P97 behavioral threshold levels.

Table 7
Cross-Paradigm Structural Invariance Results

Environment	Threshold	$S \subset B$	$S \subset IF$	$S \subset RF$	RCI_B	RCI_IF	Interpretation
WTMC2021	P95	True	True [†]	True [†]	0.158	—	Confirmed [†]
NF-ToN-IoT-v2	P95	True	True	True	0.003	0.038	Confirmed (full scale)
NF-ToN-IoT-v2	P97	True	True	True	0.003	0.074	Confirmed (full scale)
NF-CSE-CIC-IDS2018-v2	P95	True	True	False	0.000	0.000	Structural misalignment
NF-CSE-CIC-IDS2018-v2	P97	True	True	False	0.000	0.001	Structural misalignment
NF-UNSW-NB15-v2	P95	True	—	—	—	—	Epistemically meaningless
UNSW-NB15	P95	True	—	—	—	—	Epistemically meaningless

[†]Cross-paradigm invariance for WTMC2021 across RF, IF, XGB, and OCSVM was established in Albluwi et al. [8]; True values reflect results reported therein. RCI_IF value for WTMC2021 is not reproduced here; refer to the companion study for full cross-paradigm metrics.

In WTMC2021 and NF-ToN-IoT-v2, the subset relationship $S \subset B \cap RF \cap IF$ holds consistently across both behavioral threshold levels. For NF-ToN-IoT-v2, this result was obtained on the complete 16,940,496-row dataset, establishing that the invariance property is not an artifact of the stratified sampling procedure used in preliminary analysis. The low RCI values observed in NF-ToN-IoT-v2 (RCI_B = 0.003) reflect the behavioral channel's low selectivity in this environment — B flags 4,297 entities including all 31 attackers — rather than weak structural-behavioral alignment. The structural channel's role as a high-precision filter operating within a broader behavioral detection set is preserved.

In NF-CSE-CIC-IDS2018-v2, $S \subset B$ holds trivially: the single entity in S is a benign infrastructure node embedded within the large behavioral detection set by coincidence rather than by structural-behavioral corroboration. The same trivial inclusion is observed for IF, as the benign infrastructure entity is embedded within the broad anomaly set generated by IF under low-PI conditions. $S \subset RF$ fails because

the supervised Random Forest correctly identifies the entity as benign, producing a cross-channel disagreement that reflects structural misalignment rather than invariance breakdown. This result is theoretically significant: it demonstrates that the invariance property, when observed in a low-PI environment, carries no epistemological value as evidence of attack-selective corroboration.

In the homogeneous collapse environments, S encompasses a large proportion of the entity population, rendering subset relationships trivially satisfiable and epistemically uninformative.

5.6 Hybrid Detection Performance Summary

Table 8 presents consolidated hybrid detection results across all environments under the applicable threshold mechanism.

Table 8

Hybrid Detection Performance — All Environments

Environment	Threshold	Config	Flagged	TP	FP	Precision	Recall	Lift
WTMC2021	P95	Hybrid	3	3	0	1.000	0.375	2,383×
NF-ToN-IoT-v2	Gap ($\tau = 214$)	Hybrid	11	10	1	0.909	0.323	677×
NF-CSE-CIC-IDS2018-v2	Gap ($\tau = 2525$)	Hybrid	1	0	1	0.000	0.000	0×
NF-UNSW-NB15-v2	P95	Hybrid	—	—	—	0.286	—	—
UNSW-NB15	P95	Hybrid	—	—	—	0.211	—	—

The zero-precision result in NF-CSE-CIC-IDS2018-v2 is not a consequence of threshold under-selection. As demonstrated in § 5.4, alternative threshold strategies — including P95, P97, absolute thresholds of 500, and a Top-11 matched selection — consistently failed to recover an attack-selective structural tail, with maximum observed precision of 0.091 under any threshold configuration. The zero-precision outcome under gap-based thresholding therefore reflects the fundamental PI insufficiency of the environment rather than a methodological artifact.

Hybrid performance ranges from Precision = 1.000 in the high-GSD, high-PI reference environment to Precision = 0.000 in the low-PI environment, with population-expansion collapse recovery yielding Precision = 0.909 under gap-based thresholding. The GSD and PI characterization established in § 5.1 is consistent with this performance gradient across all five environments.

5.7 PI Empirical Contrast

The most direct empirical evidence for PI as an operationally necessary validity condition is provided by the methodologically matched comparison between NF-ToN-IoT-v2 and NF-CSE-CIC-IDS2018-v2. Both environments are drawn from the same source dataset (NF-UQ-NIDS-v2), processed using identical

graph construction and behavioral evaluation pipelines, and evaluated under the same gap-based threshold methodology. Table 9 summarizes the contrast.

Table 9
PI Empirical Contrast: NF-ToN-IoT-v2 vs. NF-CSE-CIC-IDS2018-v2

Property	NF-ToN-IoT-v2	NF-CSE-CIC-IDS2018-v2
Flow records	16,940,496	18,893,708
Entities	23,079	224,568
Gini	0.271	0.812
TC	Yes (Pop-Exp)	No
MGR	12.47×	1.19×
Top-20 attackers	18/20 (90%)	1/20 (5%)
First attacker rank	1	7
Tail Purity (post-hoc)	90.9%*	2.6%
Hybrid Precision	0.909	0.000
Hybrid Lift	677×	0×
$S \subset RF$	True	False

*Under gap-based threshold.

The First attacker rank indicator provides an intuitive descriptive measure of attacker prominence placement within the structural tail. High-PI environments exhibit attacker entities occupying the highest structural ranks, whereas low-PI environments exhibit structural tails dominated by benign infrastructure entities prior to the appearance of attacker nodes.

NF-CSE-CIC-IDS2018-v2 appears structurally more favorable than NF-ToN-IoT-v2 under conventional aggregate GSD indicators, exhibiting a higher Gini coefficient, no threshold collapse, and a moderate P95/Median ratio. Yet its structural channel precision is zero, compared to 0.909 in NF-ToN-IoT-v2. The difference is explained by PI: in NF-ToN-IoT-v2, the structurally prominent region is strongly attacker-enriched (MGR = 12.47×, Top-20 attackers = 18/20); in NF-CSE-CIC-IDS2018-v2, the structurally dominant entities are benign infrastructure nodes (MGR = 1.19×, Top-20 attackers = 1/20, first attacker at structural rank 7). This contrast supports PI as an empirically necessary validity dimension for structural channel utility within the evaluated environments, one that is not captured by GSD indicators alone.

5.8 Topology Validity Taxonomy

The empirical results across five environments establish a four-case topology validity taxonomy characterizing the conditions under which graph-based structural intrusion reasoning succeeds, recovers, fails, or becomes structurally impossible. Table 10 presents this taxonomy with empirical instantiation.

Table 10
Topology Validity Taxonomy for Structural Intrusion Reasoning

Case	GSD	PI	TC Type	Representative Environment	Hybrid Precision	Structural Utility
I	High	High	None	WTMC2021	1.000	Full
II	Mixed	High	Pop-Exp	NF-ToN-IoT-v2	0.909*	Recoverable
III	Medium-High	Low	None	NF-CSE-CIC-IDS2018-v2	0.000	None
IV	Low	Low	Homogeneous	UNSW-NB15, NF-UNSW-NB15-v2	0.211 / 0.286	Degenerate

*Under topology-adaptive gap-based thresholding.

The taxonomy's central implication is that Case III — high GSD with low PI — cannot be identified through aggregate inequality metrics alone, requiring the complementary PI diagnostics introduced in § 3.3. A practitioner relying exclusively on GSD assessment would incorrectly classify NF-CSE-CIC-IDS2018-v2 as a structurally suitable deployment environment, when in fact its structural channel provides no attack-selective filtering utility. The pre-deployment diagnostic protocol in § 3.7 is designed to prevent this misclassification. The empirical findings reported here constitute the observational foundation for that protocol, and their broader implications for graph-based intrusion reasoning are discussed in § 6.

The taxonomy in Table 10 characterizes static topological validity across benchmark environments with fixed entity populations. However, production deployments may not preserve a single static GSD-PI profile over time — entity populations grow continuously, traffic patterns shift across hours and days, and external events may produce non-stationary topological changes. The following section evaluates whether GSD-PI remains interpretable under temporal growth, visitor persistence patterns, and event-driven traffic shifts in a real production web environment.

5.9 Real-World Temporal Validation of GSD-PI in Production Web Traffic

5.9.1 Motivation and Environment Description

The experimental environments evaluated in § 4–§ 5 represent network intrusion detection datasets generated under controlled or semi-controlled conditions. While these environments provide the statistical properties required for rigorous framework validation, they do not expose the temporal

dynamics characteristic of production deployments — environments in which the entity population grows continuously, traffic patterns shift across hours and days, and external events produce non-stationary topological changes. This section extends the GSD-PI validity framework to a live production environment, examining whether the static validity conditions established in § 3 remain interpretable when topology evolves over a one-month observation window.

The evaluation environment is a large-scale Arabic-language religious content repository serving millions of monthly requests, operated by one of the authors under whose administrative authority all log collection was performed. Traffic data was collected continuously over a 30-day period from the server's SSL access logs, yielding 23,846,928 HTTPS request records after excluding internal automated processes. The resulting dataset comprises 961,200 unique source entities — approximately 50 times the largest entity population in the benchmark evaluation suite — spanning legitimate human visitors, authorized search engine crawlers, unauthorized automated scrapers, and adversarial actors. All analyses were performed locally; no individual IP addresses or user-identifiable information are reported in this study.

Labeling. Entity-level maliciousness labels were derived from AbuseIPDB, a community-driven IP reputation service, using a confidence threshold of ≥ 25 to classify entities as MALICIOUS. AbuseIPDB-derived labels are used throughout as external reputation-based indicators rather than definitive ground truth; the inherent limitations of community-sourced reputation scoring are acknowledged, and findings are interpreted accordingly. Authorized search engine crawlers were classified as BENIGN through known IP range matching prior to reputation lookup. A total of 1,171 IP reputation records were obtained from AbuseIPDB and matched against the production entity table. Within the full-month entity population, 355 source entities were classified as MALICIOUS and 960,845 as BENIGN, yielding an effective base rate of 0.037%.

5.9.2 Aggregate GSD-PI Characterization

Table 11 presents the aggregate GSD and PI diagnostic indicators for the production environment across the full one-month observation window, alongside the benchmark environments from § 5 for reference.

Table 11
GSD-PI Indicators: Production Web vs. Benchmark Environments

Environment	Entities	Gini	P95/Med	TC	MGR	Tail Purity	GSD	PI
WTMC2021	19,060	0.799	47,144×	No	> 500×	100%	High	Very High
NF-ToN-IoT-v2	23,079	0.271	1.0×	Yes (Pop-Exp)	12.47×	90.9%*	Mixed	High*
NF-CSE-CIC-IDS2018-v2	224,568	0.812	22×	No	1.19×	2.6%	Medium	Low
Production web (30 days)	961,200	0.762	3.2×	No	1.39×	0.7%	Medium	Low

*Under gap-based threshold.

The production environment exhibits Medium GSD (Gini = 0.762) and Low PI (Tail Purity = 0.7%), placing it structurally analogous to NF-CSE-CIC-IDS2018-v2 — an environment in which high aggregate structural inequality coexists with benign-infrastructure-dominated structural prominence. This finding is consistent with the GSD-PI dissociation established in § 5.1 and demonstrated directly in § 5.7: high Gini does not imply attack-selective structural prominence. Under P95 thresholding, the structural channel achieves Recall = 1.000 across all 355 labeled MALICIOUS entities, but Precision = 0.007, confirming structural tail domination by legitimate high-activity entities. The Top-50 analysis yields Precision = 0.680 and Lift = 1,841×, indicating substantial attacker-enrichment of the extreme structural elite despite the overall low-PI characterization.

5.9.3 Temporal Dynamics of PI: Anchor-Based Multi-Window Analysis

Benchmark-oriented GSD-PI assessment treats topological validity as fixed within a specified observation window. The 30-day production dataset enables examination of whether this holds across observation windows of different durations. Table 12 presents GSD-PI indicators over anchor-based forward windows of 1, 3, 7, 14, and 30 days. Entity counts in Table 12 reflect anchor-window populations and should not be interpreted as identical to the full-month aggregate population in Table 11.

Table 12
Anchor-Based Multi-Window Structural Validity Indicators

Window	Entities	Gini	P95/Med	MGR	Tail Purity	Precision@Top50	Lift@Top50
1 day	56,479	0.733	10.5×	2.41×	7.7%	0.380	81.3×
3 days	128,722	0.758	6.3×	2.40×	3.9%	0.540	254.6×
7 days	325,411	0.765	6.0×	1.71×	1.4%	0.600	685.1×
14 days	572,614	0.731	4.3×	1.92×	1.2%	0.580	935.5×
30 days	589,010	0.729	4.3×	1.98×	1.2%	0.600	995.5×

Two opposing temporal trends are observed simultaneously, constituting structural bifurcation: Tail Purity decreases monotonically from 7.7% at one day to 1.2% at 30 days, while Precision@Top50 increases from 0.380 to 0.600 and stabilizes after seven days. The broader structural tail (P95) is progressively contaminated as legitimate entities accumulate unique destination paths through normal repeated usage. The extreme structural elite (Top-50) remains more attacker-enriched over longer windows, suggesting that a subset of malicious or unauthorized automated entities accumulates structural prominence at a rate sufficient to remain distinguishable from the broader contaminated tail.

Spearman correlation analysis across 349 two-hour windows confirms the structural bifurcation at sub-daily timescales. Gini correlates moderately with Tail Purity ($\rho = 0.503$, $p < 0.001$), indicating that windows with higher aggregate structural inequality correspond to higher tail attack-selectivity. MGR, however, shows no significant correlation with Tail Purity ($\rho = 0.057$, $p = 0.29$) and a weak negative correlation with Precision@Top50 ($\rho = -0.189$, $p < 0.001$). This dissociation indicates that local structural isolation does not predict attack-selectivity when structural prominence can be achieved through intensive legitimate browsing behavior — consistent with the PI failure mode identified in § 5.4.

5.9.4 Dual Contamination Mechanisms

To directly test whether Tail Purity degradation is caused by legitimate entity accumulation, the 961,200 source entities were classified by active-day count — the number of distinct calendar days on which each entity appeared in the access logs.

Table 13
Tail Contamination by Visitor Persistence Category

Visitor Type	Count	% of Population	In Tail (P95)	% of Tail	Attacker% in Tail
1-day	787,129	81.9%	17,611	33.1%	0.0%
2–3 days	129,743	13.5%	12,200	23.0%	0.2%
4–7 days	27,044	2.8%	7,756	14.6%	0.5%
8–14 days	5,855	0.6%	4,270	8.0%	0.3%
15 + days	11,429	1.2%	11,298	21.3%	2.3%

The structural tail is contaminated by two empirically distinct mechanisms in approximately equal proportions. The first is session-level intensity: single-visit entities constitute 33.1% of the structural tail despite an attacker rate of 0.0%, reflecting users who browse 13 or more unique pages within a single session. The second is temporal accumulation: entities active across 15 or more days constitute 21.3% of the structural tail with an attacker rate of 2.3%, reflecting progressive accumulation of moderate daily activity into a structurally significant footprint. Top-200 analysis reveals that entities active 4–7 days achieve average out-degrees of 36,559 — exceeding the 15,774 average of persistent 15+-day entities — further complicating attack-selective thresholding.

5.9.5 Event-Driven Structural Collapse

A particularly instructive case emerges from a single-day traffic surge on May 26, 2026, when unique visitor count reached 80,666 — an 87% increase above the monthly daily average of 50,282 — temporally associated with a recurring annual observance of high significance to the site's primary user community. The site operator reports consistent server overload events during similar seasonal peaks throughout the year, providing operational corroboration that calendar-predictable collective user behavior produces recurring topology-disruption events in content repositories serving culturally homogeneous communities.

Table 14
Event-Driven Structural Collapse: Before, During, and After

Period	Unique IPs	Gini	MGR	Tail Purity	Precision@Top50
Day - 1 (Before)	49,275	0.658	3.71×	13.3%	0.480
Day 0 (Event)	80,666	0.545	1.60×	5.8%	0.220
Day + 1 (After)	47,072	0.710	2.02×	10.3%	0.420

Four structural discriminability indicators declined simultaneously while unique IP volume increased sharply: Gini from 0.658 to 0.545, MGR from 3.71× to 1.60×, Tail Purity from 13.3% to 5.8%, and Precision@Top50 from 0.480 to 0.220. Full recovery was observed within 24 hours. This pattern is consistent with an event-driven topology perturbation temporally associated with a known external

traffic surge, establishing event-driven Threshold Collapse as a third empirically distinct form of structural validity degradation. Unlike homogeneous collapse (absent structural signal) or population-expansion collapse (percentile threshold degeneracy), event-driven collapse is temporary, recoverable, and temporally associated with an identifiable external cause.

5.9.6 Implications for GSD-PI Assessment

The production validation extends GSD-PI from a static cross-dataset diagnostic framework into a temporally aware deployment assessment protocol. It demonstrates that PI is not a fixed environmental property in production web traffic: broad-tail discriminability degrades under persistence-driven and session-intensity contamination, while elite-level structural discrimination may remain useful for prioritizing high-risk entities.

Three operational extensions to the § 3.7 protocol follow. First, pre-deployment GSD-PI assessment must be supplemented with an observation-window specification. Short windows (1–3 days) preserve higher Tail Purity and are appropriate for broad structural screening where wider candidate coverage is operationally preferred; longer windows (7+ days) produce higher Precision@Top50 and are appropriate for identifying structurally persistent threat actors. Second, event-driven structural collapse should be recognized as a distinct third form of Threshold Collapse. In environments serving culturally or thematically homogeneous communities, such collapses are calendar-predictable and should be anticipated rather than treated as anomalies. Third, Tail Purity computed over long observation windows may understate short-window PI in production environments; short-window sub-sampling is recommended when precise PI characterization is operationally required.

6. Discussion

6.1 Structural Validity as a Prior Condition for Graph-Based Reasoning

The central finding of this study is that graph-based structural intrusion reasoning is not a universally applicable detection paradigm but a topology-conditioned one. The empirical results across five heterogeneous environments demonstrate that the effectiveness of the structural channel is determined primarily by the topological properties of the deployment environment — specifically, by whether the environment satisfies the GSD and PI validity conditions — rather than by the sophistication of the detection model or the comprehensiveness of the behavioral pipeline.

This framing represents a departure from the dominant evaluation paradigm in graph-based IDS research, which treats the network topology as a fixed and implicitly suitable input and focuses analytical attention on model architecture, feature engineering, and classification performance. The results reported here suggest that this focus is misplaced when deployed in environments that do not structurally support graph-based discrimination. In such environments, model improvement cannot recover a structural signal that the topology does not contain. Recognizing this constraint shifts the

diagnostic question from "how well does the model perform?" to "does this environment support the type of reasoning the model requires?" — a prior question whose answer conditions the interpretability of all downstream performance metrics.

The pre-deployment diagnostic protocol introduced in § 3.7 operationalizes this prior question as a practical assessment procedure executable from unlabeled traffic. Its value lies not in replacing model evaluation but in establishing the conditions under which model evaluation results are epistemically meaningful. A detection system achieving high cross-channel agreement (RCI) in an environment that fails the GSD-PI conditions may be producing convergent noise rather than convergent detection — a distinction that aggregate performance metrics cannot reveal but that the diagnostic protocol identifies before deployment.

6.2 The GSD-PI Dissociation: Central Theoretical Contribution

The most theoretically significant finding of this study is the empirical dissociation between structural differentiation and attack-selective prominence, most clearly instantiated by NF-CSE-CIC-IDS2018-v2.

The dissociation arises because GSD measures aggregate distributional inequality without distinguishing between inequality driven by legitimate high-connectivity infrastructure and inequality driven by anomalous attacker behavior. In enterprise network environments, server farms, gateways, domain controllers, and network management systems routinely exhibit extremely high out-degrees — not because they are malicious, but because they provide services to large numbers of clients. When such infrastructure is present, the structurally prominent region of the out-degree distribution is populated by a mixture of benign and malicious entities, and the structural filter loses attack-selectivity regardless of the degree of distributional inequality.

This finding has direct implications for the design and evaluation of graph-based IDS. Prior work establishing that source out-degree or centrality measures correlate with malicious behavior has primarily been validated in laboratory environments — including CIC-IDS2017 and its derivatives — where benign traffic is generated by a small number of workstations with limited connectivity and attacker traffic stands out structurally against this constrained background. The CIC-IDS2018 full-scale results demonstrate that this structural separability does not generalize to richer enterprise environments where legitimate infrastructure occupies the same structural region as potential attackers. Practitioners deploying centrality-based structural detection in production enterprise environments should treat PI assessment as a prerequisite, not an afterthought.

6.3 Population-Expansion Collapse and the Limits of Percentile Thresholding

The population-expansion collapse phenomenon identified in NF-ToN-IoT-v2 reveals a distinct failure mode for percentile-based structural thresholding in large-scale IoT deployments. As IoT networks scale

to tens of thousands of devices — the majority of which maintain minimal connectivity to a small number of local gateways — the P95 threshold degenerates to out-degree = 1, flagging the entire entity population and rendering the structural channel operationally equivalent to no filtering at all.

The critical insight is that this degeneracy is not caused by the absence of structural signal but by the mismatch between the threshold mechanism and the topological regime. The attacker entities in NF-ToN-IoT-v2 exhibit out-degrees of 200–5,000, representing connectivity patterns that are structurally extreme by any reasonable measure. The P95 threshold fails to identify them not because they are structurally inconspicuous, but because the threshold is pulled down by the overwhelming majority of peripheral devices. Gap-based thresholding, which identifies the natural structural break in the distribution rather than an arbitrary percentile of the full population, recovers a high-precision structural set (Precision = 0.909, Lift = 677 $\times$) in an environment where percentile thresholding produces Lift $\approx$ 1.0 $\times$.

This finding suggests that percentile-based structural thresholds, which are standard in the graph-based IDS literature, are poorly suited for IoT and large-scale heterogeneous network deployments where the entity population is dominated by peripheral low-degree devices. Topology-adaptive thresholding mechanisms — derived from natural distributional discontinuities rather than arbitrary percentile cutoffs — are more appropriate for these environments and should be considered in the design of structural detection pipelines intended for operational deployment at scale. The maximum gap ratio diagnostic provides a label-free mechanism for identifying environments where this adaptation is necessary.

6.4 Temporal Validity and Production Web Traffic

The production validation in § 5.9 extends the static GSD-PI framework in two theoretically significant directions. First, it demonstrates that PI is not a fixed environmental property but a temporally varying one: the same deployment environment exhibits substantially different PI values at different observation-window lengths, ranging from 7.7% Tail Purity at one day to 1.2% at 30 days. This temporal variability is explained by two contamination mechanisms — session-level intensity and temporal accumulation — that progressively populate the structural tail with legitimate entities as the observation window extends.

Second, the event-driven structural collapse documented in § 5.9.5 establishes a third empirical type of Threshold Collapse not present in the benchmark environments. Unlike homogeneous collapse and population-expansion collapse, which reflect fixed topological properties, event-driven collapse is temporary, externally caused, and in some deployment contexts calendar-predictable. This distinction carries operational significance: a diagnostic system that responds to event-driven collapse by permanently reconfiguring the structural channel would be incorrect; one that recognizes the transient nature of the collapse and schedules a reassessment after the event period would be appropriate. The GSD-PI pre-deployment protocol should therefore be extended to distinguish between permanent and transient validity degradation in production environments.

6.5 Cross-Paradigm Invariance as a Topology-Conditioned Property

The cross-paradigm invariance results address whether structural prominence is a stable property across behavioral detection paradigms or an artifact of a specific detection mechanism. The results from WTMC2021 and NF-ToN-IoT-v2 indicate the former. Structurally prominent entities under sufficient GSD and PI conditions are consistently embedded within behavioral detection sets of threshold-based, Isolation Forest, and Random Forest paradigms — methods that differ substantially in statistical assumptions, training requirements, and operational characteristics.

The theoretical interpretation is that structural prominence under sufficient PI reflects a genuine behavioral characteristic: extreme communication breadth that manifests as anomalous across multiple representational channels simultaneously. Cross-paradigm invariance is therefore an emergent property of the underlying behavioral phenomenon rather than a methodological artifact. This is supported by the contrasting result in NF-CSE-CIC-IDS2018-v2, where the structurally dominant entity is a benign infrastructure node — correctly identified as benign by the supervised Random Forest — breaking the invariance relationship and confirming that the structural signal points in the wrong direction. Cross-paradigm invariance therefore provides supporting evidence for the GSD–PI interpretation: its breakdown in a low-PI environment is consistent with structural misalignment rather than serving as an independent test of PI sufficiency.

The low RCI values observed in NF-ToN-IoT-v2 — despite confirmed subset relationships — reflect the behavioral channel's limited selectivity when B flags thousands of entities in a large IoT population, making Jaccard similarity necessarily small even when $S \subset B$ holds exactly. This reinforces the recommendation that RCI be interpreted in conjunction with GSD and PI assessment rather than as a standalone corroboration metric.

6.6 Relationship to the Companion Study

This paper extends the companion study in three distinct directions. First, it formalizes the operational conditions under which the structural-behavioral convergence demonstrated in the companion study is epistemically reliable — conditions that the companion study demonstrated empirically but did not systematically characterize. Second, it identifies and distinguishes the two failure modes — PI insufficiency and Threshold Collapse — that bound the scope of the companion study's findings. Third, it provides full-scale empirical validation of cross-paradigm structural invariance on data an order of magnitude larger than the companion study's primary evaluation environment, establishing that the observed invariance is not an artifact of the controlled laboratory conditions or the sampling procedure.

Importantly, the results of this paper do not contradict or weaken the companion study's findings. WTMC2021 satisfies the GSD and PI conditions established here; the companion study's results are therefore epistemically reliable within their scope. What this paper establishes is the boundary of that scope: the companion study's structural-behavioral convergence finding generalizes to environments satisfying sufficient GSD and PI, and does not generalize to environments that fail either condition.

The relationship between the two papers can be summarized concisely: the companion study demonstrated the phenomenon; this paper characterizes when it holds, why it holds, and under what conditions it loses validity. This relationship is explicitly anticipated in the companion study's future work directions, which identified multi-environment validation across topologically heterogeneous and IoT-specific environments as a defined next step. The present paper delivers this extension while contributing the GSD-PI theoretical framework that characterizes when and why the companion study's findings generalize across heterogeneous deployment environments.

6.7 Limitations

Several limitations bound the generalizability of the findings reported here.

Evaluation scope. The topology validity taxonomy is established across five environments. While these environments span qualitatively distinct topological regimes and include full-scale analysis of two datasets exceeding 16 million flow records each, the classification boundaries — including the MGR thresholds used to distinguish homogeneous from population-expansion collapse — are derived from this evaluation suite and may not transfer universally to topological configurations not represented here. Additional evaluation across cloud-native, encrypted-traffic, and operational enterprise environments would strengthen the generalizability claims.

Temporal granularity. Although § 5.9 evaluates GSD-PI across multiple observation windows, the structural representation within each window remains a static aggregated graph. Temporal graph models that preserve event order and evolving graph states remain a defined direction for future work.

PI operationalization. The label-free PI diagnostics — Tail Density and Maximum Gap Ratio — provide indirect evidence of attack-selective structural prominence without directly measuring it. The connection between high MGR and attack-selective prominence is empirically supported across the evaluated environments but is not theoretically guaranteed: a high MGR in a novel deployment environment may reflect the isolation of legitimate infrastructure rather than attacker prominence. Post-hoc Tail Purity validation, where feasible, provides the most direct confirmation of PI sufficiency.

Gap threshold calibration. The gap-based threshold τ_{gap} is derived from a data-driven procedure that selects the most restrictive significant structural break producing a non-trivial detection set. The minimum gap ratio ($\gamma_{\text{min}} = 3.0\times$) and minimum set size ($k_{\text{min}} = 2$) parameters governing this procedure were selected empirically from the evaluated environments and may require recalibration for environments with substantially different structural characteristics. The sensitivity of detection outcomes to these parameters across a wider range of topological configurations has not been systematically evaluated. The selected γ_{min} value separates the empirically observed homogeneous and population-expansion regimes within the evaluated environments, but should not be interpreted as a theoretically universal constant; environments exhibiting MGR values in the intermediate range (approximately $2\times$ – $5\times$) likely represent a partially characterized boundary regime requiring further calibration and evaluation across additional topological configurations.

7. Conclusion

This paper reframed graph-based intrusion detection as a topology-conditioned validity problem. Its central argument is that before evaluating how effectively a detection model exploits graph structure, one must first examine whether the network environment itself provides structural evidence capable of supporting intrusion reasoning. Structural differentiation alone is insufficient: a topology may exhibit substantial inequality while its most prominent entities are legitimate high-connectivity infrastructure, whereas apparent threshold failure may occur even when a recoverable attack-enriched structural tail remains present.

The empirical results demonstrate this distinction directly. Across the evaluated environments, structural differentiation and attack-selective prominence exhibit different operational regimes. The most informative contrast is between NF-ToN-IoT-v2 and NF-CSE-CIC-IDS2018-v2. Despite exhibiting substantially greater aggregate structural inequality, NF-CSE-CIC-IDS2018-v2 produces Hybrid Precision = 0.000 under gap-based structural selection, whereas NF-ToN-IoT-v2 achieves Precision = 0.909 and Lift $\approx 677\times$ after recovery from population-expansion collapse. This GSD-PI dissociation provides empirical evidence that aggregate structural inequality alone cannot establish structural detection validity.

The topology validity taxonomy further distinguishes four observed regimes: full structural utility, recoverable structural utility, infrastructure-dominated structural failure, and structurally degenerate environments. Threshold Collapse provides an additional diagnostic distinction between genuine absence of recoverable structural differentiation and percentile-threshold failure caused by population expansion. Cross-paradigm analysis supports this interpretation by showing that structural-behavioral agreement is informative only when the structural set itself is meaningfully isolated; subset agreement in low-PI or degenerate environments can instead reflect convergent non-discrimination.

Production validation over 23.8 million HTTPS requests extends the framework beyond static benchmark conditions. The results show that prominence isolation is temporally sensitive: broad-tail purity deteriorates as observation windows expand, while highly ranked structural subsets can remain strongly enriched. Observation-window length therefore becomes an important operational parameter when applying topology-validity diagnostics in long-running deployments.

The resulting pre-deployment protocol operationalizes GSD, PI, Threshold Collapse assessment, and threshold recovery from unlabeled network traffic before structural detection results are interpreted. Its practical purpose is to distinguish model limitations from topology-level constraints. Without this distinction, poor structural detection performance may be incorrectly attributed to model architecture when the deployment environment itself does not provide sufficiently selective structural evidence.

The broader methodological implication is that graph-derived evidence should not be assumed valid solely because a graph representation can be constructed or because the topology exhibits statistical heterogeneity. Similar prior-validity questions may arise in other graph-based analytical domains where structural prominence is used as evidence, although such transfer requires separate empirical validation.

For graph-based intrusion detection, the present results establish a more specific conclusion: **structural reasoning is topology-conditioned, and its validity should be diagnosed before deployment rather than inferred from downstream model performance alone.**

Declarations

Author Contributions

Abdulhadi Albluwi conceived the GSD-PI validity framework, designed the pre-deployment diagnostic methodology, implemented the full experimental pipeline, conducted the graph-based structural and behavioral analyses across five evaluation environments, performed the cross-paradigm invariance validation, interpreted the results, and wrote the manuscript.

Mohamed I. Marie and Helal A. Suleiman contributed to research supervision, methodological review, and manuscript revision.

All authors reviewed and approved the final manuscript.

Funding

No funds, grants, or other support was received for conducting this study.

Competing Interests

The authors have no relevant financial or non-financial interests to disclose.

Code Availability Statement

The implementation code, GSD and PI assessment procedures, cross-paradigm invariance validation, and figure generation scripts supporting this study are openly available at:

<https://github.com/AbdalhadiB/GSD-PI-Structural-Validity-IDS>. The repository includes: (i) the complete analysis notebook (paper2_gsd_analysis.ipynb), which downloads NF-UQ-NIDS-v2 directly from Kaggle and reproduces all reported GSD and PI indicators, structural detection results, and cross-paradigm invariance findings; and (ii) the Figure 2 generation script (figure2_ccdf.py).

Data Availability Statement

The production SSL access logs analyzed in §5.9 are not publicly released as they contain operational server telemetry and potentially sensitive IP-level access metadata. Only aggregate statistics are reported in this manuscript. No individual IP addresses from the production environment are disclosed.

The datasets analysed in this study are publicly available. NF-UQ-NIDS-v2 (comprising NF-ToN-IoT-v2, NF-CSE-CIC-IDS2018-v2, and NF-UNSW-NB15-v2) is available via Kaggle at <https://www.kaggle.com/datasets/dhoogla/nfuqnidsv2>. The WTMC2021 dataset was generated using the publicly released relabelling code available at <https://github.com/GintsEngelen/WTMC2021-Code>.

The UNSW-NB15 dataset is available from the UNSW Cyber Range Lab at <https://research.unsw.edu.au/projects/unsw-nb15-dataset>.

References

1. Cybersecurity Ventures. (2025). *2025 Official Cybercrime Report*. Cybersecurity Ventures.
2. IBM Security. (2025). *Cost of a Data Breach Report 2025*. IBM.
3. Zoubir, N., & Missaoui, R. (2024). Graph neural networks for network intrusion detection. *Computers & Security*, 143, 103982. <https://doi.org/10.1016/j.cose.2024.103982>
4. Dilini, N., Sun, N., Miao, Y., & Moustafa, N. (2026). A comprehensive review on graph-based anomaly detection: Approaches for intrusion detection. *Applied Sciences*, 16(4), 1906. <https://doi.org/10.3390/app16041906>
5. Wu, C., Sun, J., Chen, J., Alazab, M., Liu, Y., & Xiang, Y. (2025). TCG-IDS: Robust network intrusion detection via temporal contrastive graph learning. *IEEE Transactions on Information Forensics and Security*, 20, 1475–1486. <https://doi.org/10.1109/TIFS.2024.3507345>
6. Farrukh, Y. A., Wali, S., Khan, I., & Bastian, N. D. (2025). XG-NID: Dual-modality network intrusion detection using a heterogeneous graph neural network and large language model. *Expert Systems with Applications*, 287, 128089. <https://doi.org/10.1016/j.eswa.2025.128089>
7. Gilliard, E., Liu, J., & Aliyu, A. A. (2024). Knowledge graph reasoning for cyber attack detection. *IET Communications*, 18, 297–308. <https://doi.org/10.1049/cmu2.12736>
8. Albluwi, A., Marie, M. I., & Suleiman, H. A. (2026). *Windowed Flow Operators for Intrusion Detection: Frozen Validation and Behavioral Corroboration Limits*. Manuscript under review at *Computer Networks*.
9. Govindarajan, S., & Muzamal, S. (2025). Advanced cloud intrusion detection framework using graph-based features, transformers, and contrastive learning. *Scientific Reports*, 15, 20511. <https://doi.org/10.1038/s41598-025-07956-w>
10. Mwiga, K., Dida, M., Maglaras, L., Mohsin, A., Janicke, H., & Sarker, I. H. (2026). Graph convolution neural network and deep Q-network optimization-based intrusion detection with explainability analysis. *Sensors*, 26(5), 1421. <https://doi.org/10.3390/s26051421>
11. Ares-Robledo, F., Rifà-Pous, H., & Clarisó, R. (2026). Graph neural networks for anomaly detection: A systematic review of dynamic temporal approaches. *Artificial Intelligence Review*, 59, 129. <https://doi.org/10.1007/s10462-026-11532-7>
12. Belcastro, L., Carlucci, C., Cosentino, C., Liò, P., & Marozzo, F. (2026). Enhancing network security using knowledge graphs and large language models for explainable threat detection. *Future Generation Computer Systems*, 176, 108160. <https://doi.org/10.1016/j.future.2025.108160>
13. Alkahtani, H., & Ahmad, H. F. (2025). Enhancing cybersecurity through autonomous knowledge graph construction by integrating heterogeneous data sources. *PeerJ Computer Science*, 11, e2768. <https://doi.org/10.7717/peerj-cs.2768>

14. Sommer, R., & Paxson, V. (2010). Outside the closed world: On using machine learning for network intrusion detection. In *2010 IEEE Symposium on Security and Privacy* (pp. 305–316). IEEE.
<https://doi.org/10.1109/SP.2010.25>
15. Sarhan, M., Layeghy, S., Gallagher, M., & Portmann, M. (2023). From zero-shot machine learning to zero-day attack detection. *International Journal of Information Security*, 22, 947–959.
<https://doi.org/10.1007/s10207-023-00676-0>
16. Koumar, J., Hynek, K., Čejka, T., & Šiška, P. (2025). CESNET-TimeSeries24: Time series dataset for network traffic anomaly detection and forecasting. *Scientific Data*, 12, 338.
<https://doi.org/10.1038/s41597-025-04603-x>
17. Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>
18. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765–4774.
<https://doi.org/10.5555/3295222.3295230>
19. .19Xu, J., Huang, L., Zhang, F., Niu, Z., Shi, K., & Li, Q. (2025). The current state of research on reputation evaluation of network nodes. *Electronics*, 14(19), 3900.
<https://doi.org/10.3390/electronics14193900>

Figures

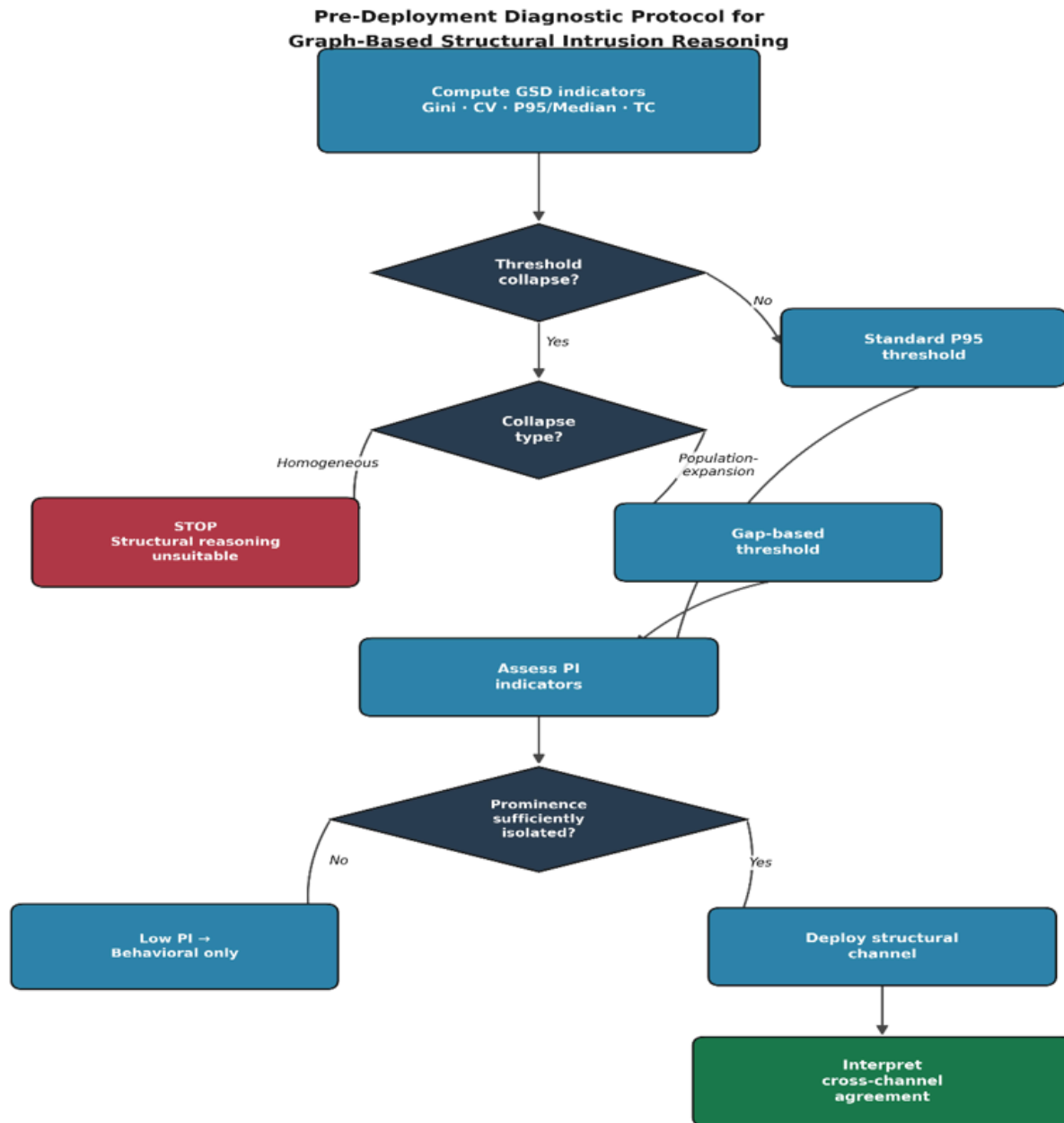


Figure 1

Pre-deployment diagnostic protocol for graph-based structural intrusion reasoning. The protocol operationalizes GSD and PI assessment as a sequential decision procedure executable from unlabeled network traffic, distinguishing threshold stability, homogeneous collapse, and recoverable population-expansion collapse before structural-channel deployment and cross-channel interpretation

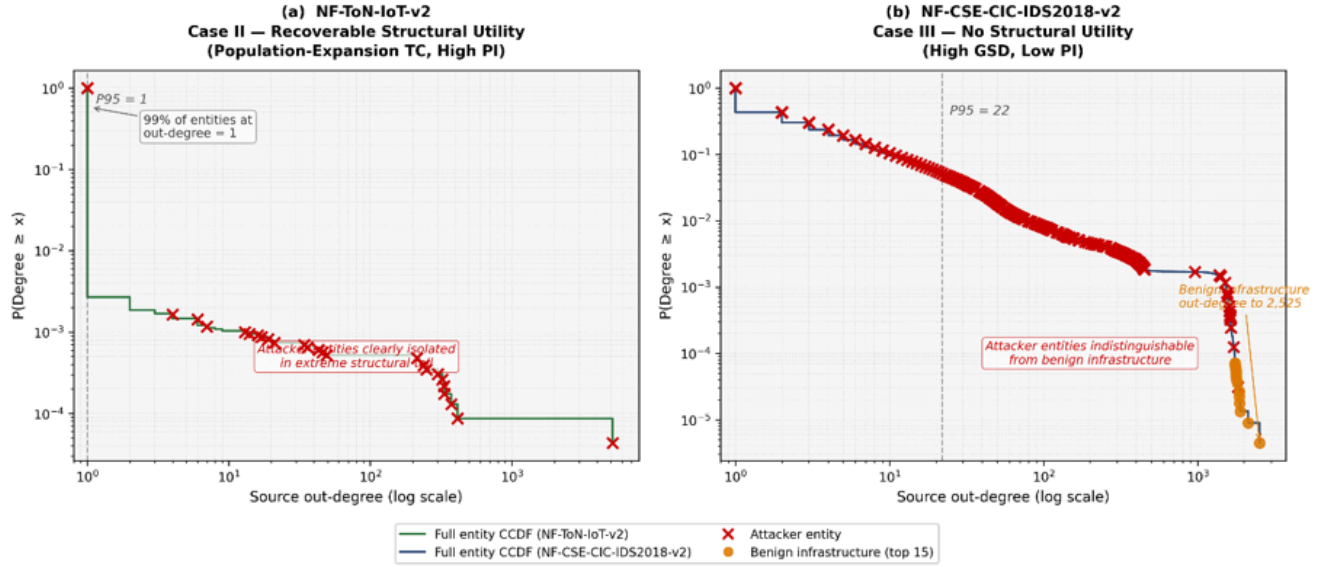


Figure 2

Source out-degree complementary cumulative distribution functions (CCDF) for NF-ToN-IoT-v2 (Case II — Recoverable Structural Utility, Population-Expansion TC, High PI) and NF-CSE-CIC-IDS2018-v2 (Case III — No Structural Utility, High GSD, Low PI). In NF-ToN-IoT-v2, attacker entities (x) are structurally isolated in the extreme tail, confirming attack-selective prominence under gap-based thresholding. In NF-CSE-CIC-IDS2018-v2, attacker entities are indistinguishable from benign high-connectivity infrastructure nodes (o), which dominate the structural tail despite high aggregate inequality. The contrast shows that recoverable structural prominence does not imply structural discriminability unless the prominent region also satisfies PI



Figure 3

Topology validity taxonomy for graph-based structural intrusion reasoning. The four-case classification is determined by joint GSD and PI assessment, with evaluated environments positioned according to their empirically confirmed structural characteristics. Case I (WTMC2021) represents full structural utility; Case II (NF-ToN-IoT-v2) establishes that population-expansion collapse is recoverable through topology-adaptive thresholding when PI is sufficient; Case III (NF-CSE-CIC-IDS2018-v2) represents the primary failure mode – structurally rich environments where the prominent region is dominated by benign infrastructure; Case IV (UNSW-NB15, NF-UNSW-NB15-v2) represents structurally degenerate environments where no discriminative signal is present