ADVERSARIAL TECHNIQUES FOR BYPASSING GRAPH NEURAL NETWORKS BASED NETWORK DEFENSE

Kartikeya Sharma
Senior Associate Information Security Engineer @ Equinix

Dr. Jun Li
Director @ Center for Cyber Security and Privacy, University of Oregon

Disclaimer

The views and opinions expressed in this presentation are my own and do not necessarily reflect the official policy or position of my employer.

According to Zayo, the average DDoS attack cost victim **$6000 per minute**

Current state of the art GNN based defenses achieve up to 99% detection accuracy

(based on open source datasets)

But there’s a problem…

State-of-the-art GNN-based DDoS detectors are fundamentally broken

→ F1 Score went down to 1% under our attacks

Our Contribution

EXPOSE

Three novel adversarial attacks that completely break state-of-the-art GNN-based Network Intusion Detection Systems (NIDS)

- Spoofed Flow Distribution (Uniform/Random)
- Spoofed Flow Distribution + Benign Injection
- Link Congestion Distribution (4 variants)

PROVE

Comprehensive experimental validation across multiple datasets and SOTA models

- 3 real-world datasets
- 2 state-of-the-art GNN models
- Catastrophic F1 score degradation

Presentation Outline

- **Background: What is GNN and how GNN-based NIDS Work?**
- **Threat Model & Attack Philosophy**
- **Three Novel Attack Methods**
- **Experimental Setup**
- **Attack Results & Impact**
- **Implications & Discussion**

Background: What is GNN and how GNN-based NIDS Work?

What are Graph Neural Networks?

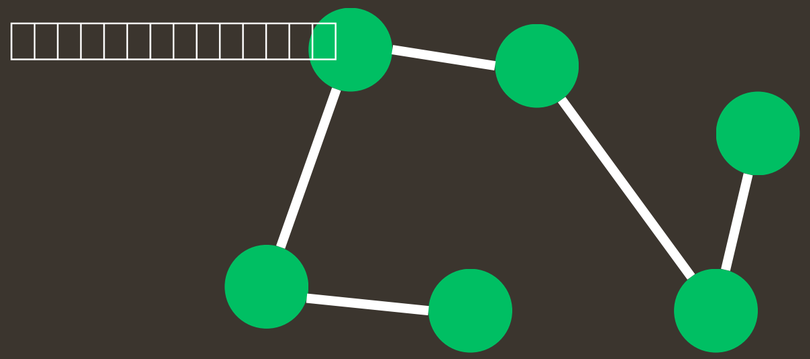
Graph Neural Networks (GNNs): Deep learning models that operate on graph-structured data

Graph Components

- **Nodes:** Entities (e.g., IPs, flows)
- **Edges:** Relationships between nodes
- **Features:** Attributes of nodes/edge

Why GNNs?

- Capture relational patterns
- Learn from network structure
- Powerful for connected data

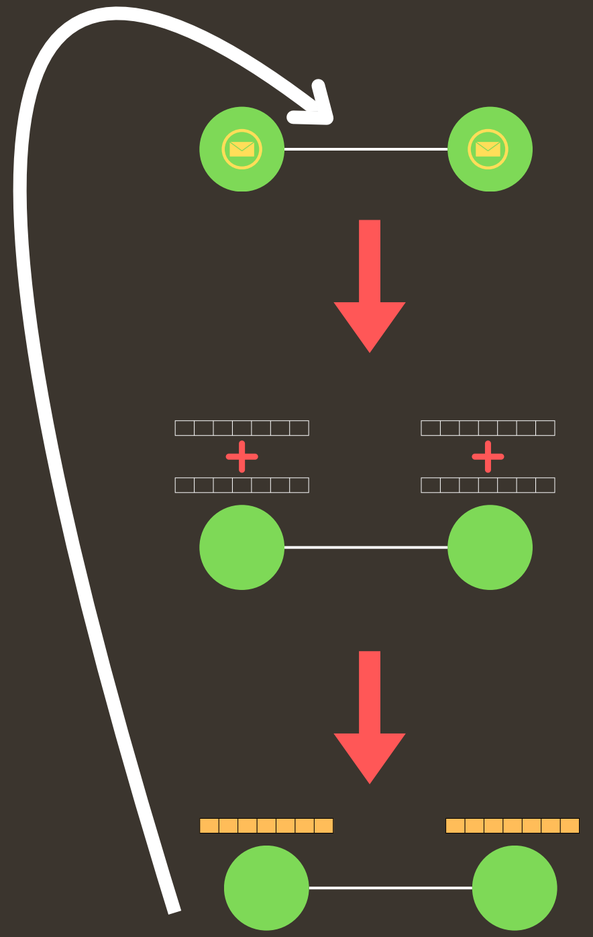


How GNNs Work: The Core Mechanism

GNNs learn rich node representations, called embeddings using **Message Passing.**

Message Passing Framework

1. **Aggregate:** Collect information from neighboring nodes
2. **Combine:** Merge neighbor info with node's own features
3. **Update:** Generate new node representation
4. **Repeat:** Multiple layers for broader context

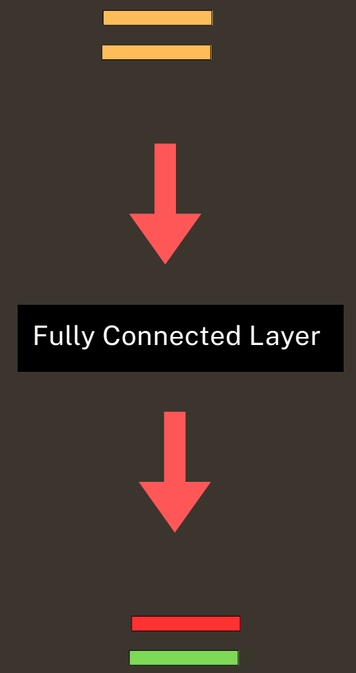


How GNNs Work: The Core Mechanism

GNNs learn rich node representations, called embeddings using **Message Passing.**

Message Passing Framework

1. **Aggregate:** Collect information from neighboring nodes
2. **Combine:** Merge neighbor info with node's own features
3. **Update:** Generate new node representation
4. **Repeat:** Multiple layers for broader context



GNNs Already Powering Industry Defenses

Vendor	Application	How GNNs Are Used
Palo Alto Networks (Unit 42)	Malicious domain discovery	Expand known IoCs across domain/IP/cert relationships
Darktrace (DIGEST)	Incident prioritization	Predict which incidents will escalate using graph dynamics
Vectra AI (GraphMDN)	Network anomaly detection	Model enterprise relationships for threat detection
Microsoft (NetVigil)	Data-center traffic monitoring	East-west anomaly detection in production clusters

GNN-based defenses are already protecting real networks which means our attacks have immediate real-world impact

Why GNNs for DDoS Detection?

Traditional ML Limitations

- Treat flows independently
- Miss relational patterns
- Ignore network topology
- Lower detection rates

GNN Advantages?

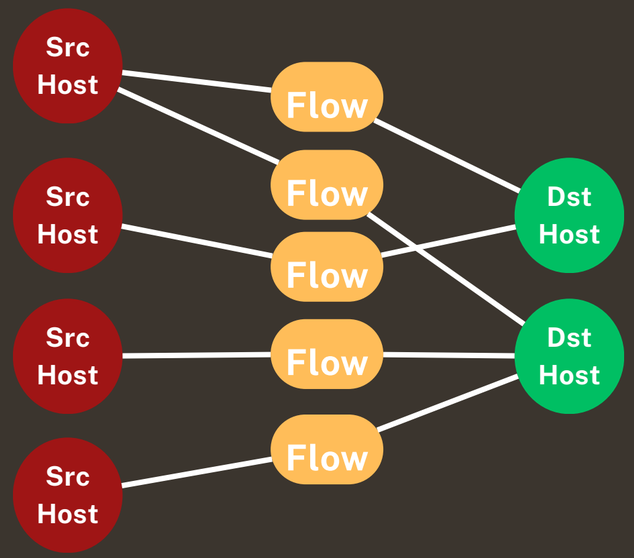
- Capture attack coordination
- Detect distributed patterns
- Leverage network structure
- 95-99% accuracy (on open-source datasets)

The Problem: This reliance on graph structure is exactly what makes GNNs vulnerable to our attacks

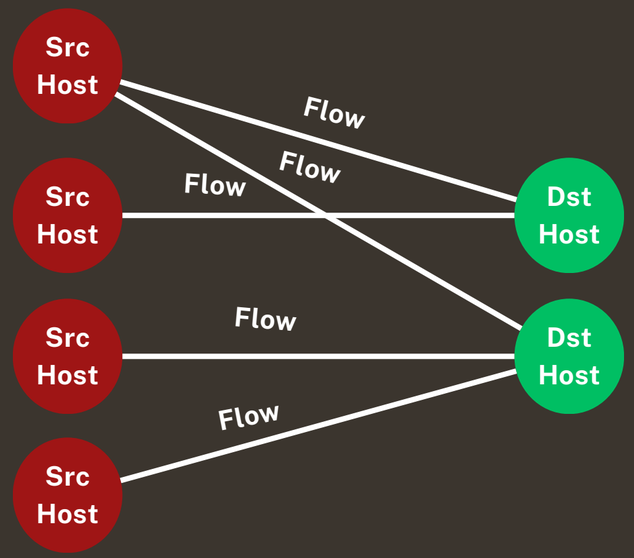
How to turn Network Flows into Graphs?

Network Flow: Aggregated packet statistics between two hosts

(SrcIP, DstIP, SrcPort, DstPort, Protocol) → Features: duration, bytes, packets, flags...

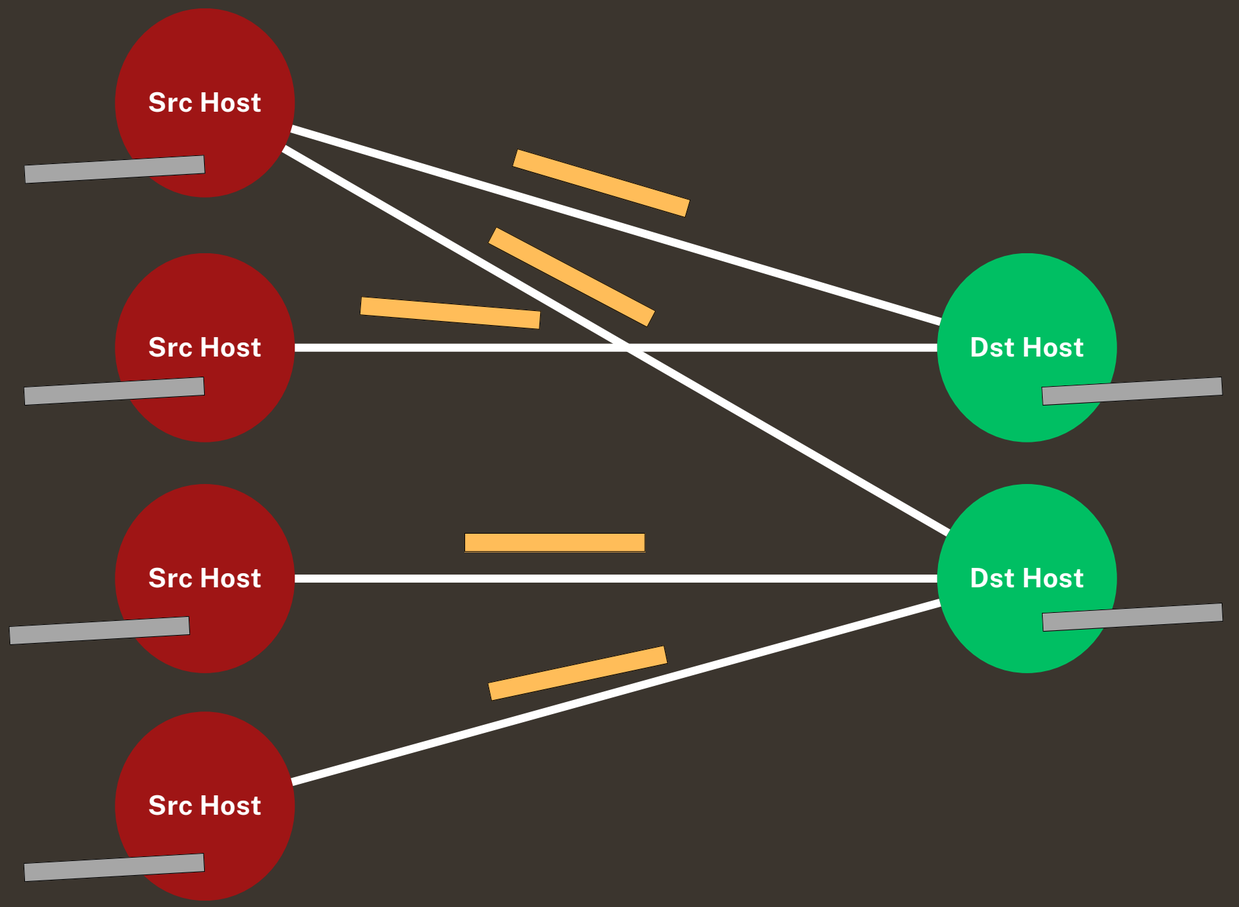


Host-Connection Graph

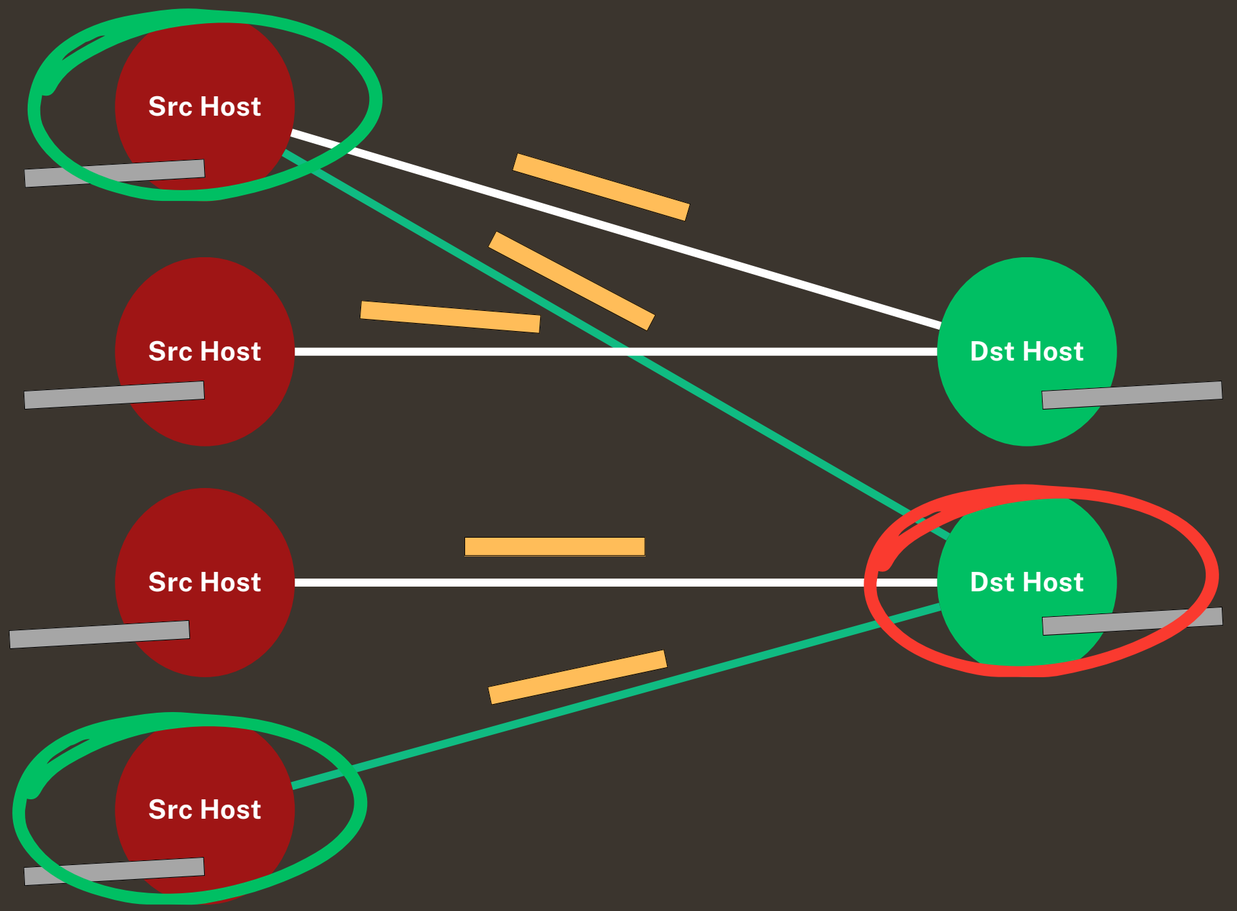


Flow Graph

SOTA Model: E-GraphSage



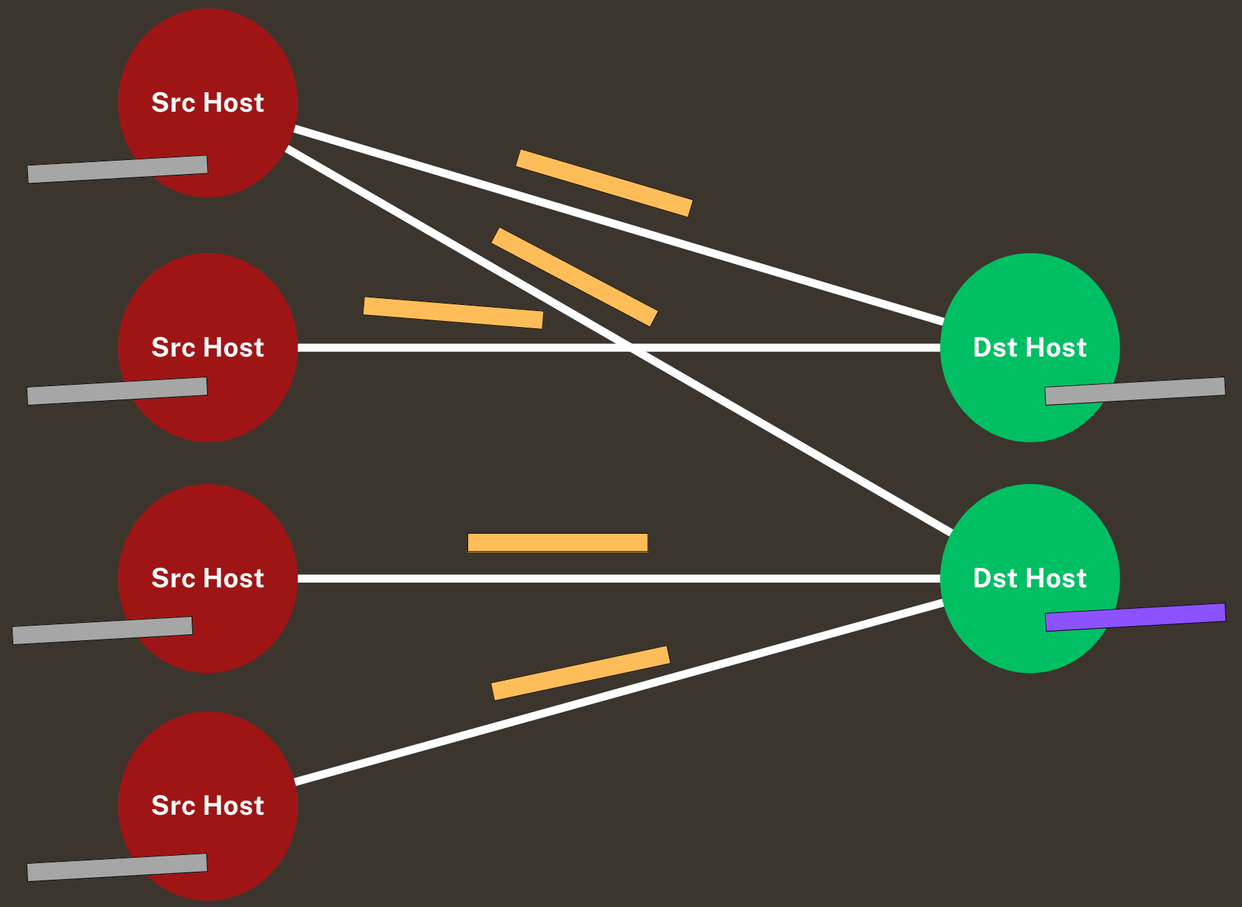
SOTA Model: E-GraphSage



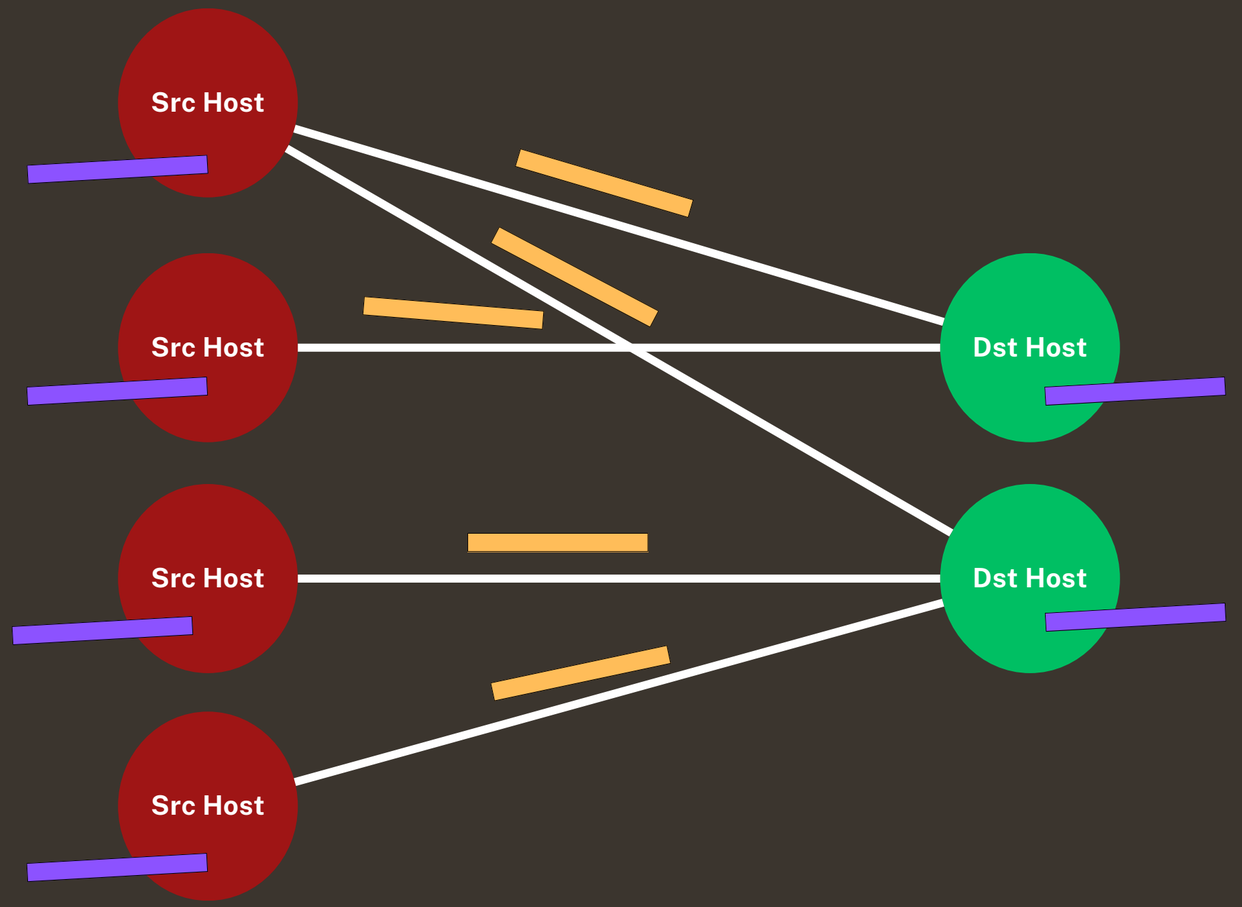
SOTA Model: E-GraphSage

SOTA Model: E-GraphSage

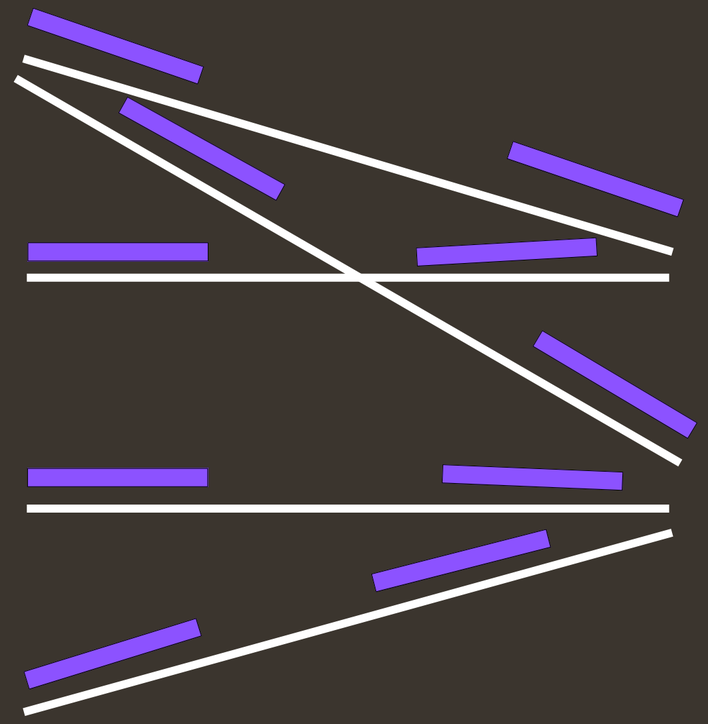
SOTA Model: E-GraphSage



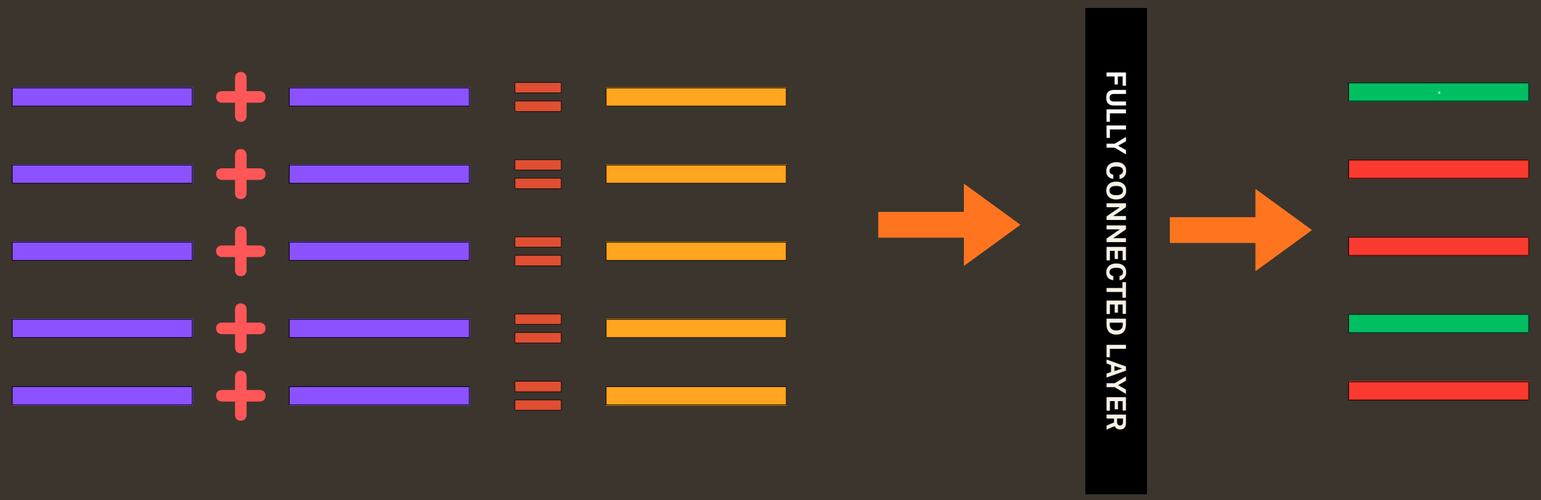
SOTA Model: E-GraphSage



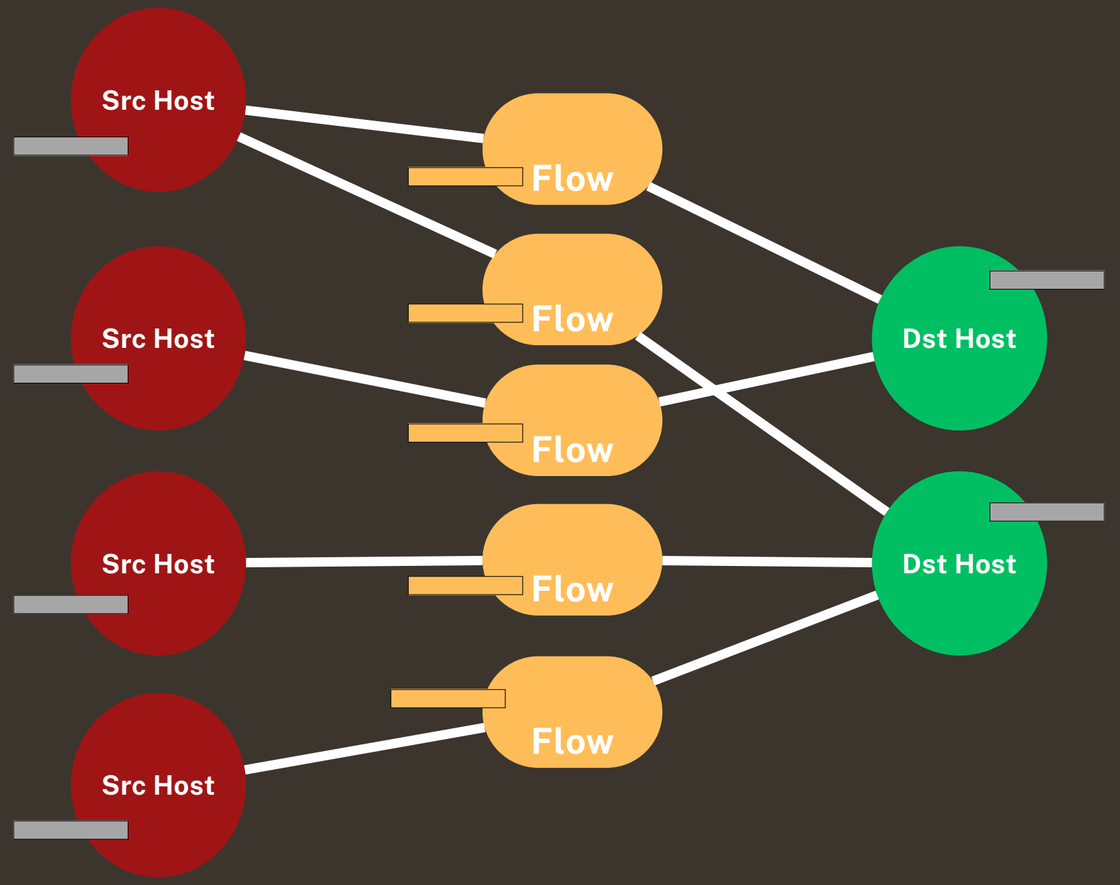
SOTA Model: E-GraphSage



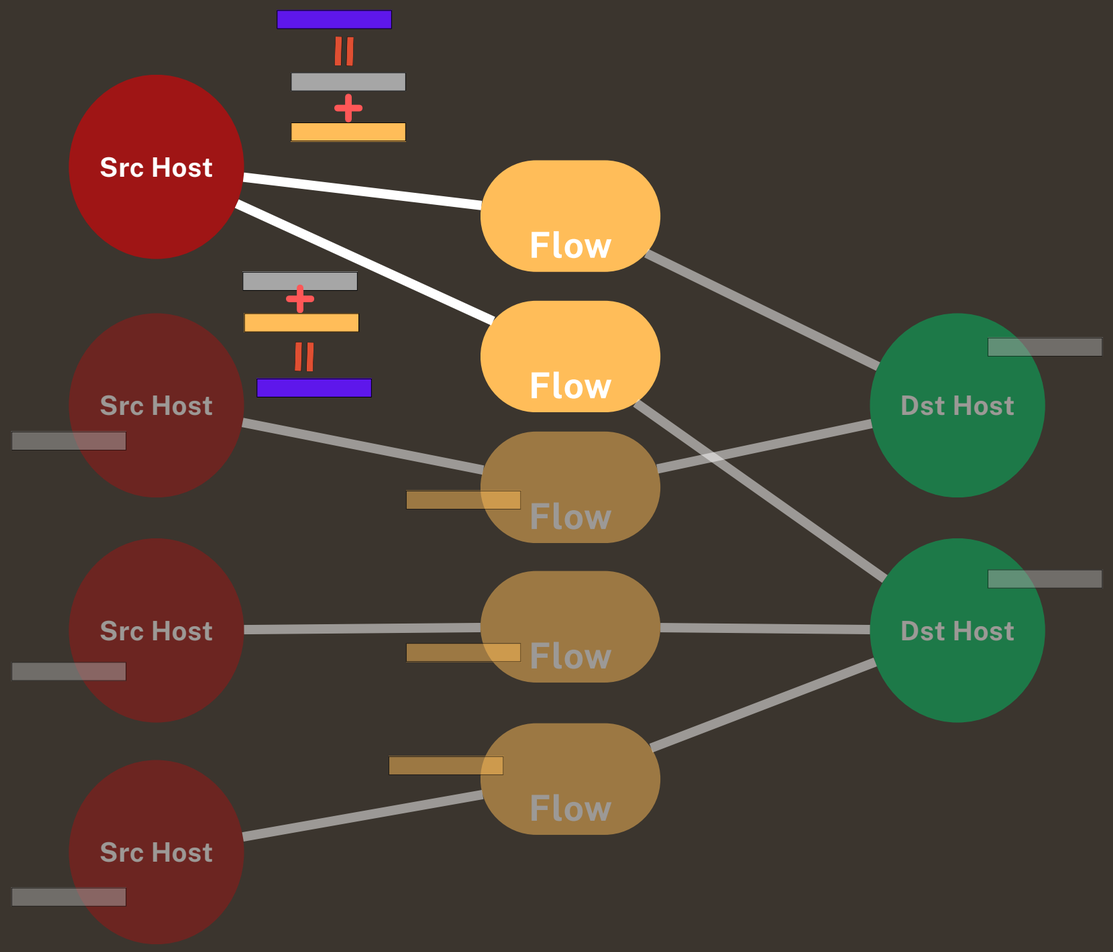
SOTA Model: E-GraphSage



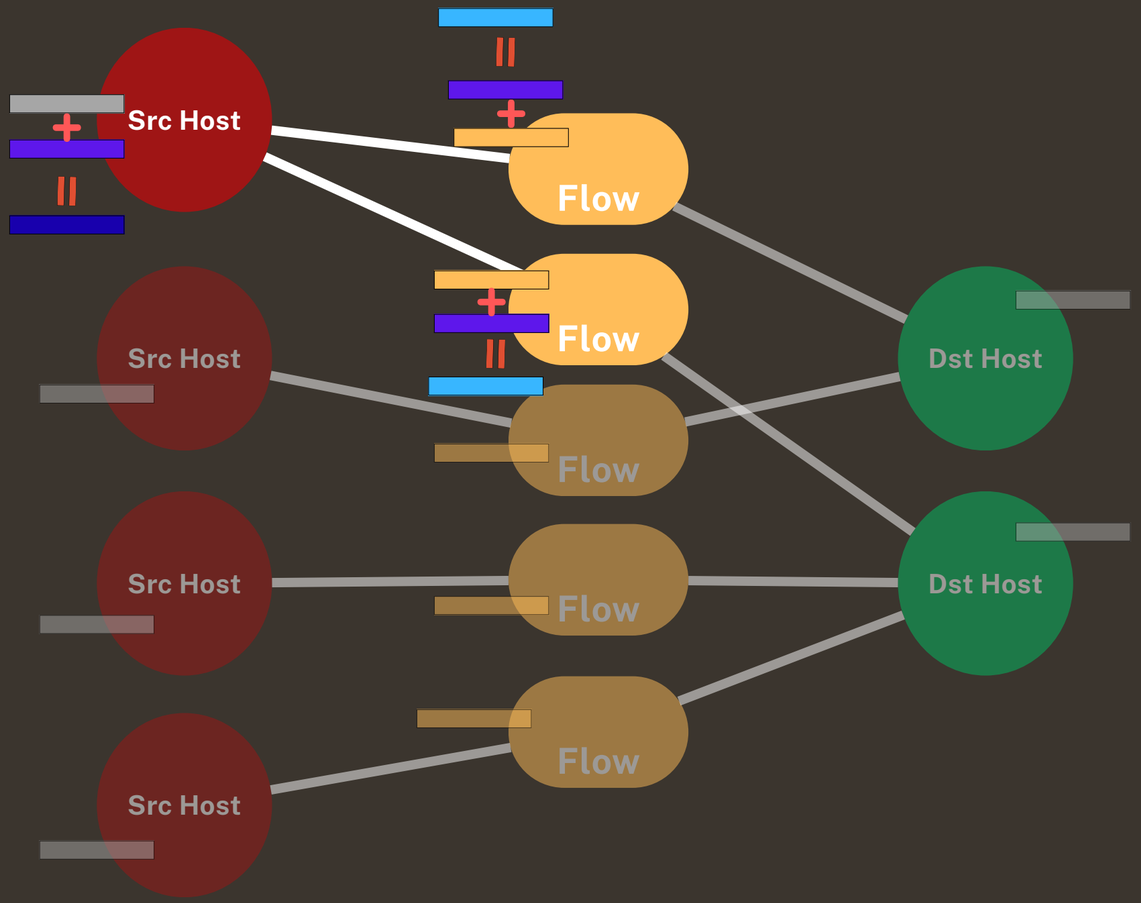
SOTA Model: GNN-RNIDS



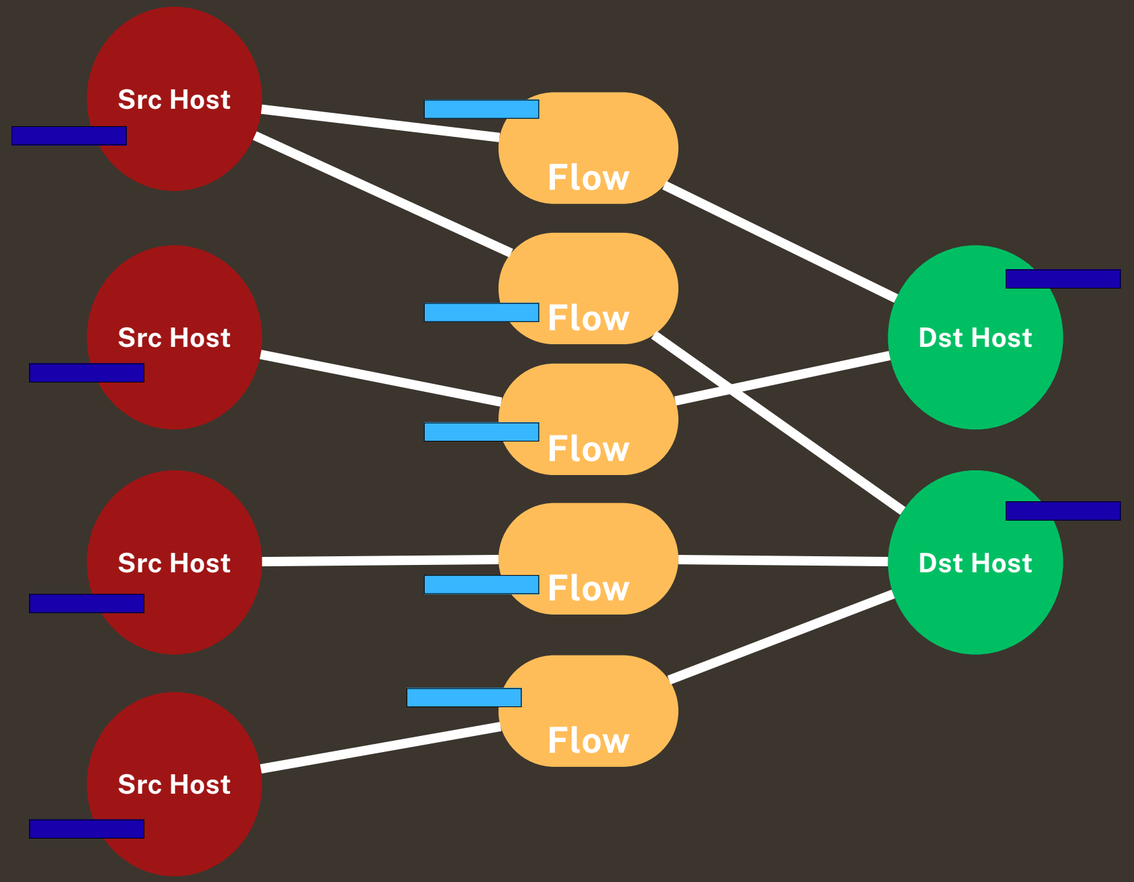
SOTA Model: GNN-RNIDS



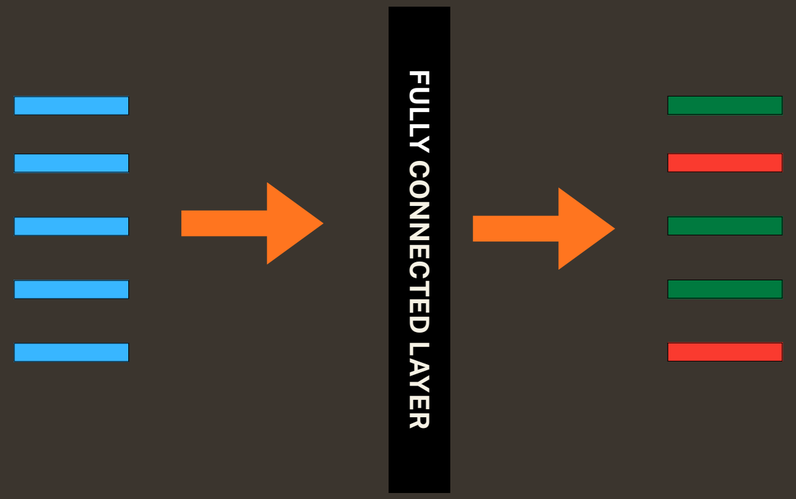
SOTA Model: GNN-RNIDS



SOTA Model: GNN-RNIDS



SOTA Model: GNN-RNIDS



The Fundamental Vulnerability

GNN based Detection Systems rely on graph topology structure derived from network metadata (IPs, ports, connections)

The Exploit:

Attackers can manipulate graph structure (IPs, connections) while preserving DDoS functionality (bandwidth consumption, service disruption)

Threat Model & Attack Philosophy

Threat Model

Adversary Goal: **Evade GNN-based detection while maintaining DDoS attack effectiveness**

Adversary Capabilities:

- Can spoof source IP addresses (standard DDoS technique)
- Can mix malicious and benign traffic
- Can redistribute attack traffic across network
- Has knowledge of target network topology

Constraints: **Must preserve attack effectiveness**

Attack Philosophy

Core Principle: Change the graph structure that GNNs depend on while maintaining DDoS attack functionality

Three Attack Categories:

Spoofed Flow Distribution

Fragment traffic across IPs

Benign Injection

Mixed behavior patterns

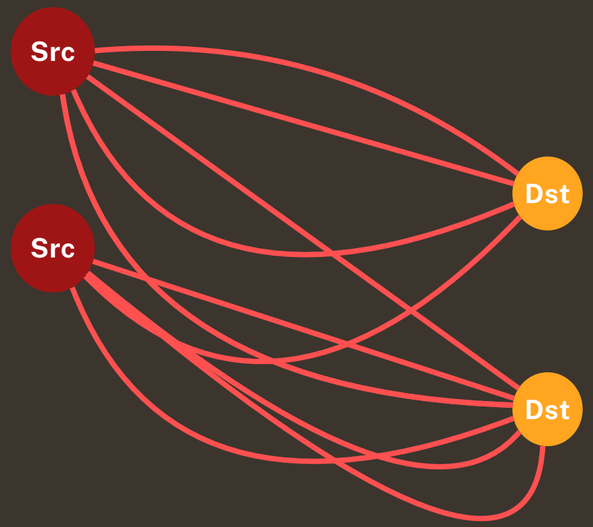
Link Congestion

Target infrastructure

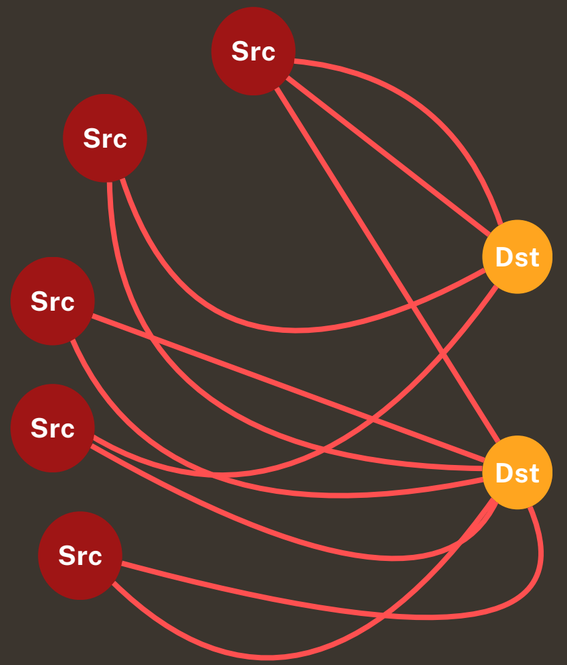
Attack 1: Spoofed Flow Distribution (SFD)

Attack Strategy

Introduce multiple spoofed malicious IPs for each original malicious IP, then redistribute traffic across these spoofed IPs to break the GNN defense.



Normal Attack



Spoofed flow distribution Attack

Attack 1: Spoofed Flow Distribution

Attack Strategy

Introduce multiple spoofed malicious IPs for each original malicious IP, then redistribute traffic across these spoofed IPs to break the GNN defense.

Uniform Distribution (USFD)

Evenly distribute flows or packets across spoofed IPs

- Flow-based: Equal flow count per spoofed IP
- Packet-based: Equal packet volume using min-heap

Random Distribution (RSFD)

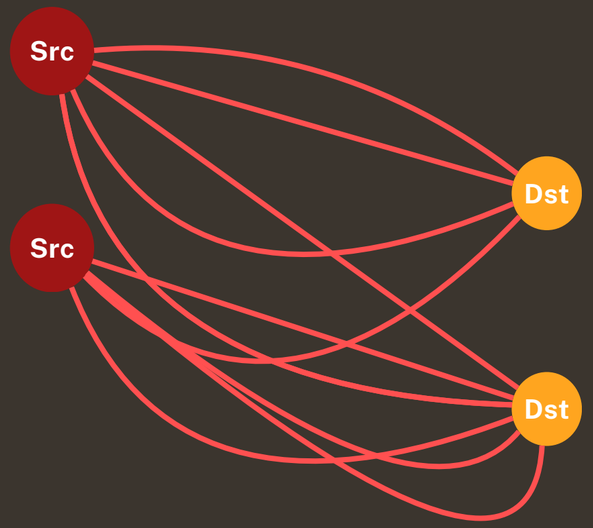
Randomly assign flows or packets to create unpredictable patterns

- Flow-based: Random assignment with guarantee of ≥1 flow per IP
- Packet-based:Random with uneven packet load distribution

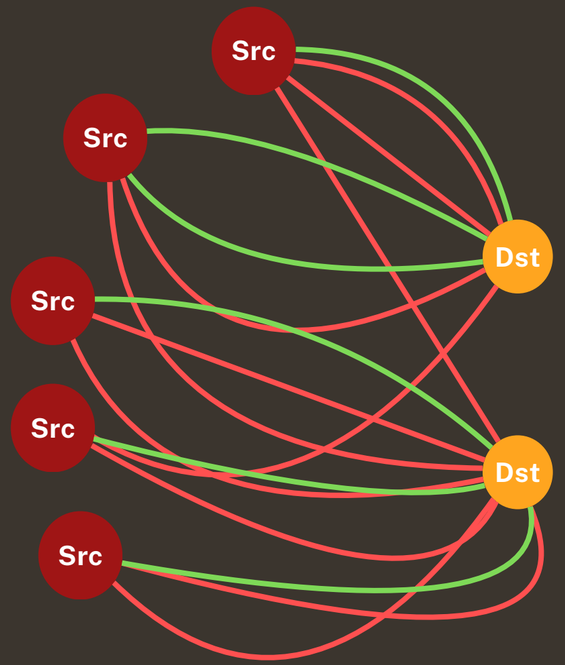
Attack 2: SFD + Benign Injection

Attack Strategy

Phase 1: Apply SFD (Uniform or Random) to distribute malicious traffic
Phase 2: Inject benign traffic from same spoofed IPs to create mixed behavior



Normal Attack

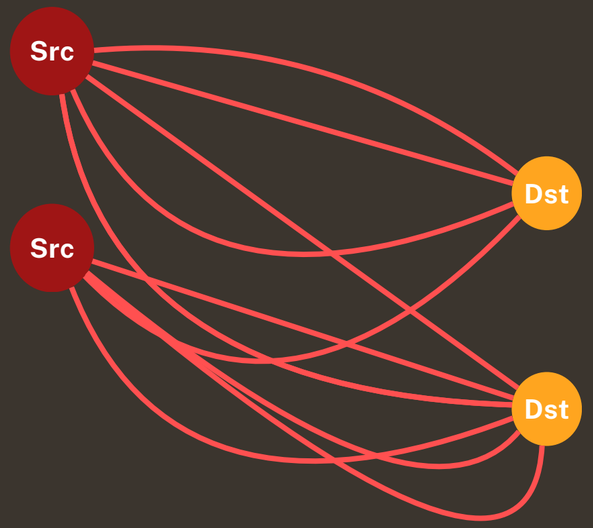


SFD + Benign Injection Attack

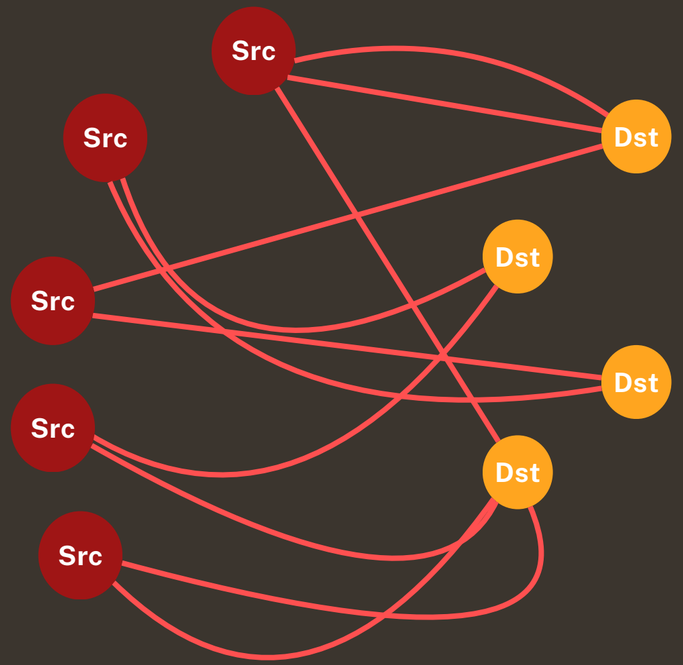
Attack 3: Link Congestion Distribution

Attack Strategy

Phase 1: Source spoofing with uniform/random distribution
Phase 2: Redirect flows to concentrated legitimate destination IPs (link targets)



Normal Attack



Link Congestion Distribution Attack

Attack 3: Link Congestion Distribution

Attack Strategy

Phase 1: Source spoofing with uniform/random distribution
Phase 2: Redirect flows to concentrated legitimate destination IPs (link targets)

Uniform Source Distribution

Even flow distribution across spoofed sources, then redirect to targets + decoys.

Random Source Distribution

Random source assignment (botnet-like), then same destination redirection pattern.

Variable Decoy Generation

Generate 5-100 decoy IPs per target subnet with mixed selection strategies. This results in unpredictable patterns

With Benign Injection

Add Phase 3: Inject 5-100 benign flows from spoofed sources to 20-100% of decoy IPs

Experimental Setup

Experimental Setup

Datasets

- **CIC IDS 2017**
  - 2.4M flows
  - 128K DDoS flows
  - 84 features
- **BCCC CIC IDS 2017**
  - 1.9M flows
  - 95K DDoS flows
  - 123 features
- **BCCC-CPacket Cloud DDoS 2024**
  - 700K flows
  - 228K DDoS flows
  - 321 features

Target Models

- **E-GraphSage**
  - Flow Graph
  - 429 citations on google scholar
- **GNN-RNIDS**
  - Host-Connection Graph
  - 125 citations on google scholar

Evaluation

- **10-fold cross-validation for baselines**
- **F1-score is used as a primary metric for attacks' evaluation**
- **Attack intensity variations**

Attack Results & Impact

Results: CIC IDS 2017 Dataset

F1 scores of Baseline and Attacks across different models

Model	Baseline	SFD Only	SFD + Benign	Link Congestion
E-GraphSAGE	99.78%	99.65%	4.60%	47.94%
GNN-RNIDS	99.17%	98.73%	61.94%	27.98%

Results: BCCC CIC IDS 2017 Dataset

F1 scores of Baseline and Attacks across different models

Model	Baseline	SFD Only	SFD + Benign	Link Congestion
E-GraphSAGE	99.99%	99.97%	0.59%	68.80%
GNN-RNIDS	99.82%	16.19%	16.12%	14.71%

Results: BCCC CPacket Cloud DDoS 2024 Dataset

F1 scores of Baseline and Attacks across different models

Model	Baseline	SFD Only	SFD + Benign	Link Congestion
E-GraphSAGE	92.61%	92.51%	45.85%	55.81%
GNN-RNIDS	95.94%	84.70%	68.84%	57.93%

Attack Effectiveness Patterns

E-GraphSAGE

Best Attack = SFD + Benign Injection

SFD alone FAILS across all datasets:

- CIC-IDS: 99.65% (no drop)
- BCCC-CIC: 99.97% (no drop)
- BCCC-CPacket: 92.51% (minimal)

SFD + Benign is devastating:

- CIC-IDS: 4.60% (-95%)
- BCCC-CIC: 0.59% (-99%)
- BCCC-CPacket: 45.85% (-51%)

GNN-RNIDS

Best Attack = Link Congestion

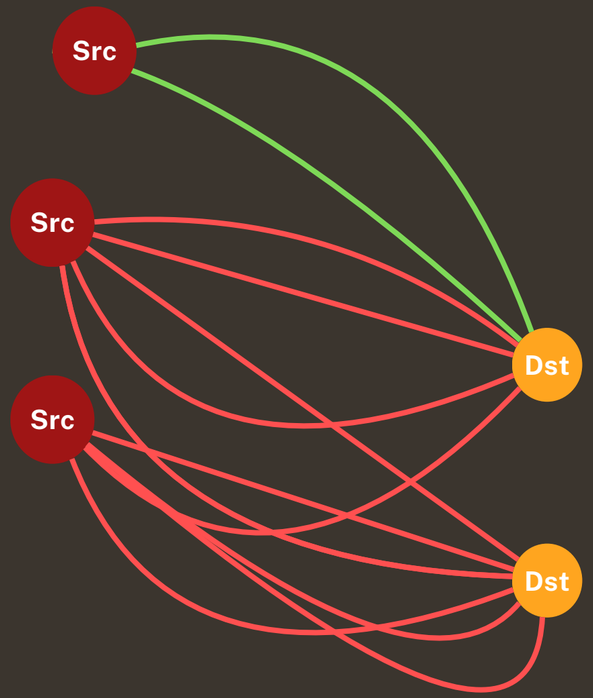
SFD alone mostly fails

- CIC-IDS: 98.73% (no drop)
- BCCC-CIC: 16.19% (-83.63%)
- BCCC-CPacket: 84.70% (minimal)

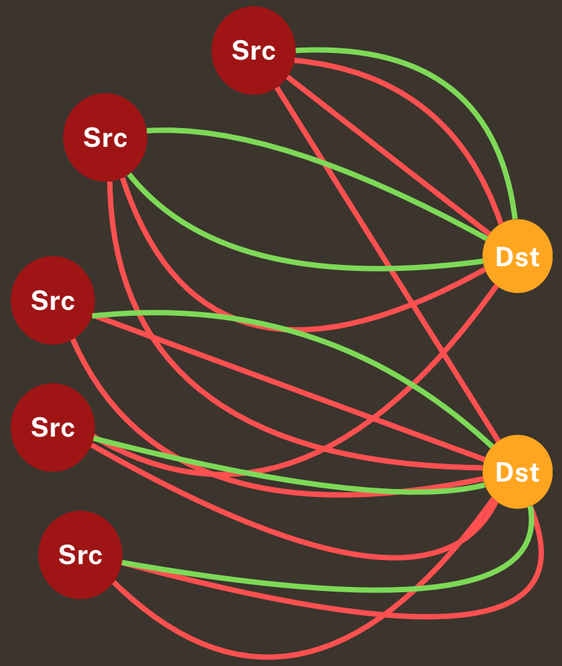
Link Congestion consistently works:

- CIC-IDS: 27.98% (-71.19%)
- BCCC-CIC: 14.71% (-85.10%)
- BCCC-CPacket: 57.93% (-38.01%)

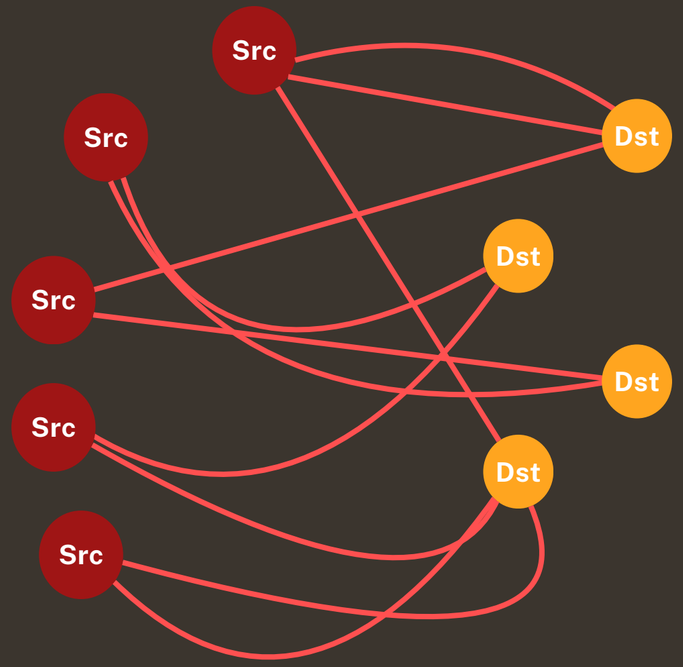
Why SFD + Benign Injection breaks E-GraphSAGE?



Normal Attack

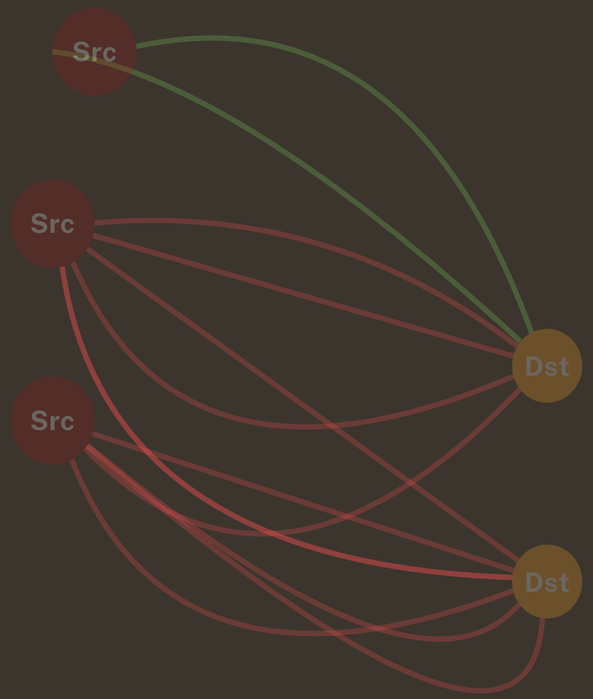


SFD + Benign Injection Attack

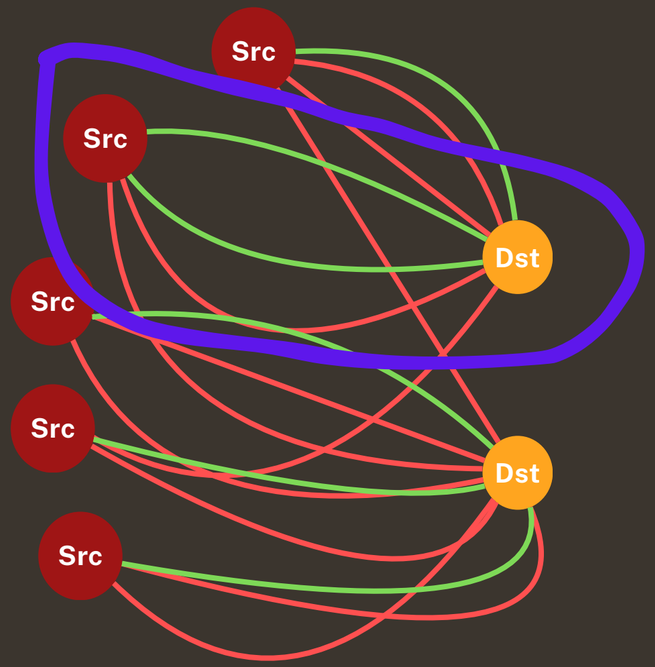


Link Congestion Distribution Attack

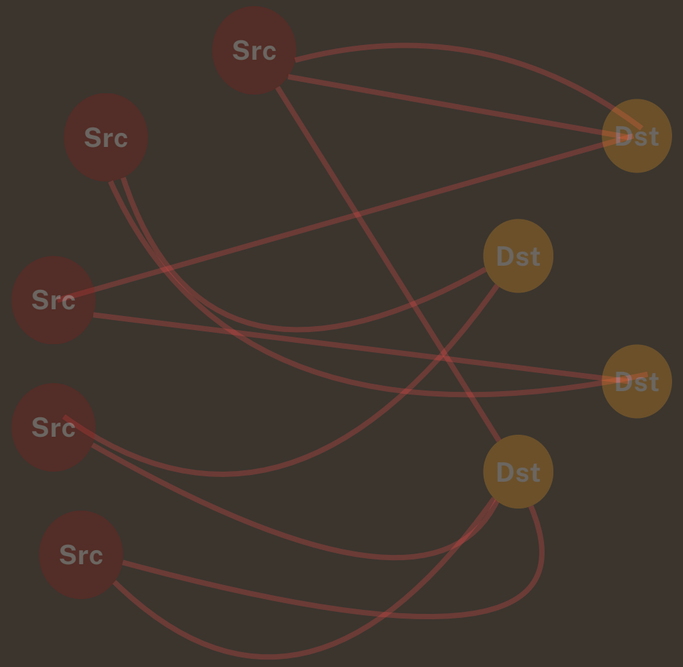
Why SFD + Benign Injection breaks E-GraphSAGE?



Normal Attack

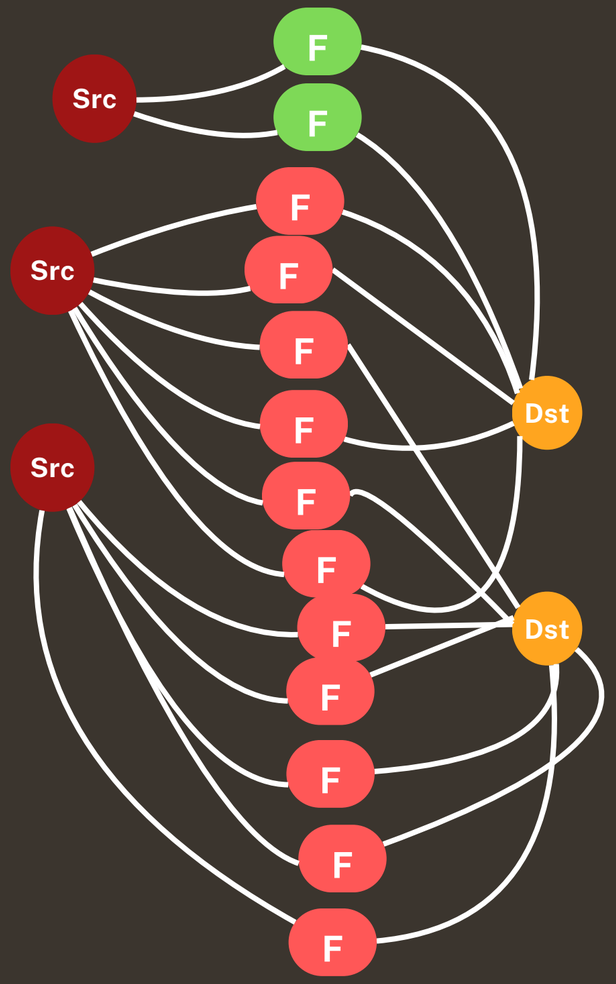


SFD + Benign Injection Attack

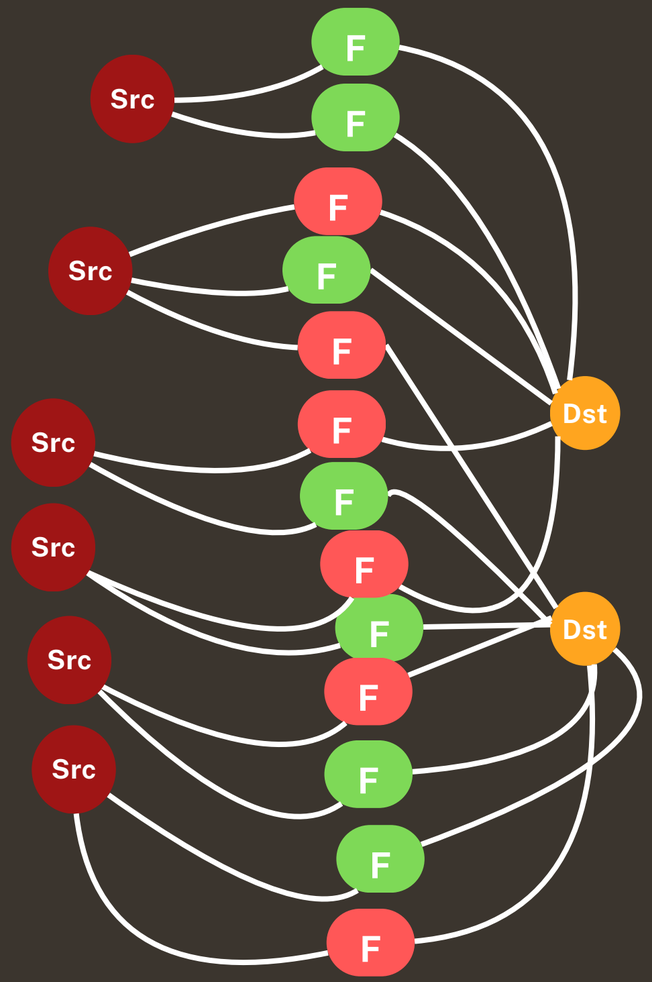


Link Congestion Distribution Attack

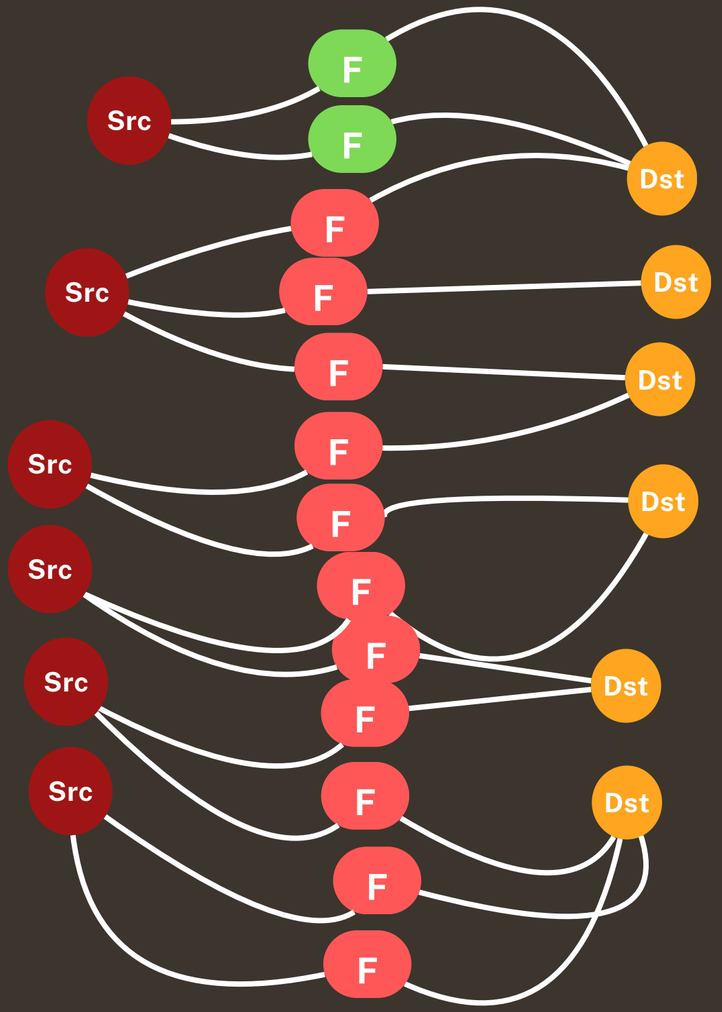
Why Link Congestion works better for GNN-RNIDS?



Normal Attack

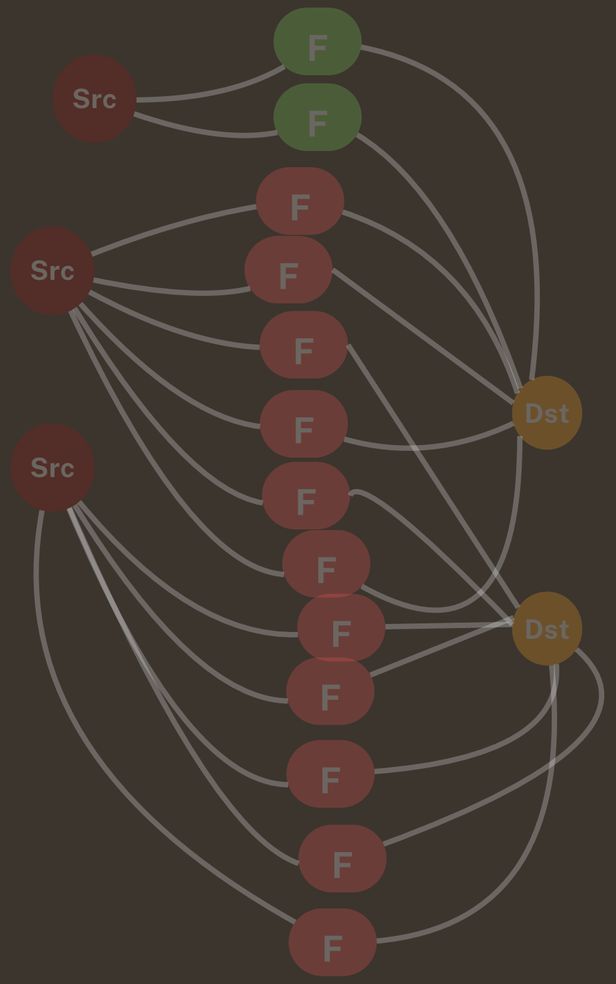


SFD + Benign Injection Attack

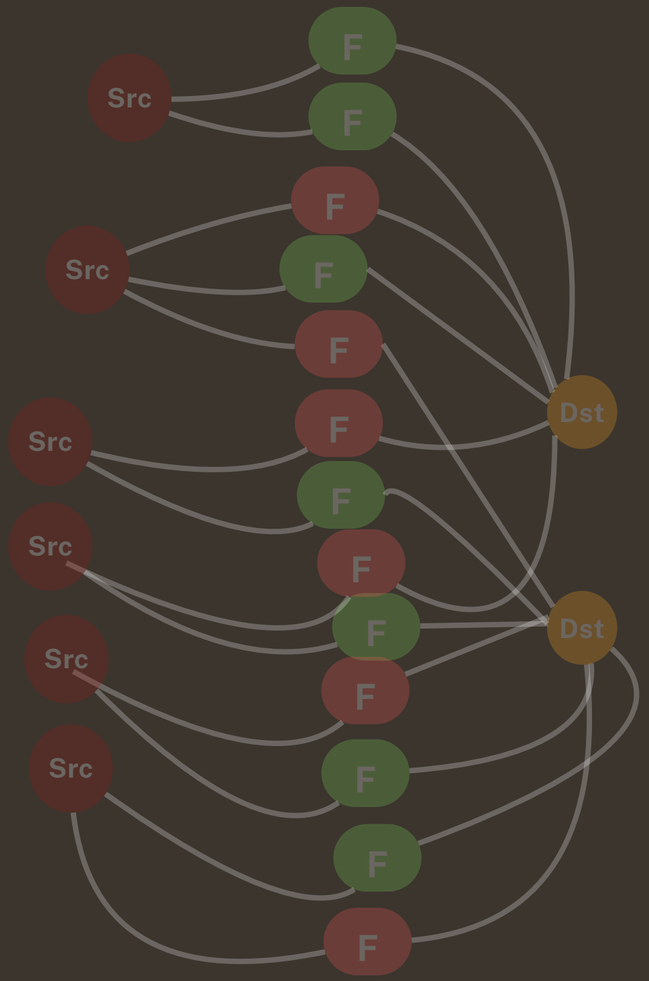


Link Congestion Distribution Attack

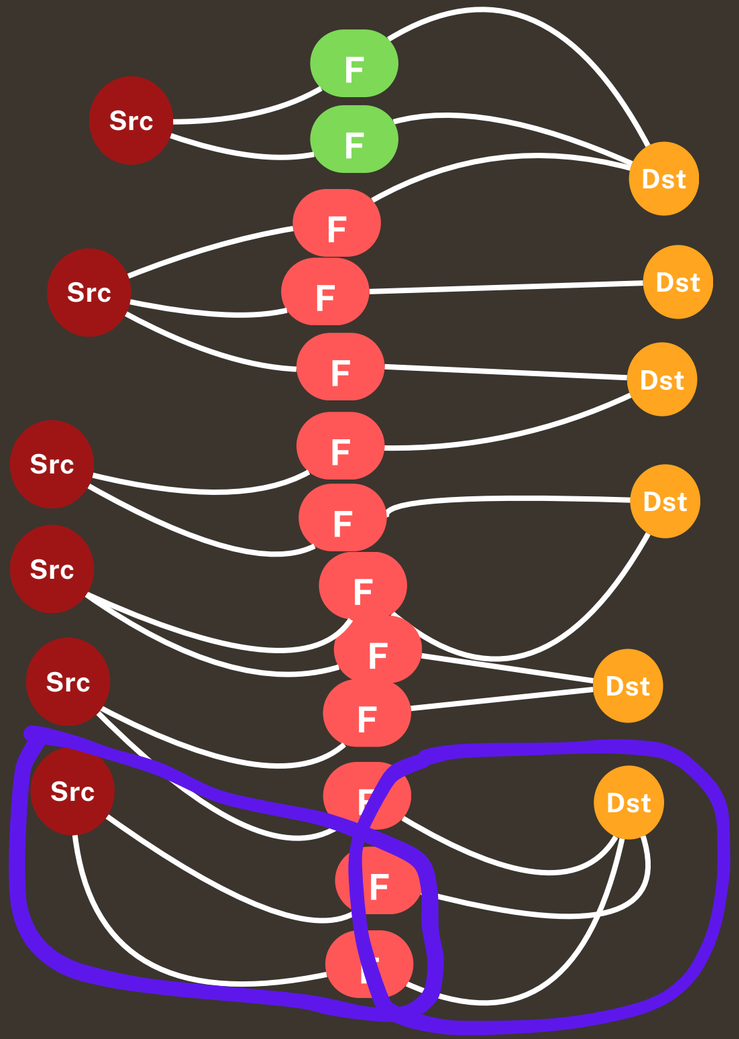
Why Link Congestion works better for GNN-RNIDS?



Normal Attack



SFD + Benign Injection Attack



Link Congestion Distribution Attack

Implications & Discussion

Real World Attack Feasibility

All attacks use standard DDoS techniques

Resource Requirements

- **IP Spoofing:** 3000-5000 IPs
- **Benign Traffic:** 45-100 flows per target
- **Link Targets:** 100 destinations
- **Decoy IPs:** 40-100 per subnet

Implementation Reality

- Compatible with existing DDoS toolkits
- No special ML or security expertise needed
- Botnet operators already have required capabilities
- Attack overhead: <5% vs standard DDoS

Critical Point

These attacks don't require new capabilities. Any adversary with basic botnet access can exploit these vulnerabilities. The attacks are practical, scalable, and deployable today.

Industry Implications

Critical: Deployed GNN-based NIDS are operationally vulnerable

Validated Threats:

- **E-GraphSAGE systems:** 50-99% f1 score drop across environments with SFD+Benign injection
- **GNN-RNIDS systems:** 40-85% f1 score drop with all attack categories
- **Even cloud environments (321 features):** 40-50% f1 score drops - protection inadequate
- **Attacks are practical:** 5000 IPs, 50-100 benign flows - achievable with botnets

Required Actions:

- **Immediate:** Test deployed GNN systems against adversarial attacks
- **Architecture:** Identify which GNN type you're using and its specific vulnerability
- **Strategy:** Consider hybrid approaches or topology-independent detection
- **Monitoring:** Watch for traffic redistribution and benign injection patterns

Key Findings Summary

1. Two state-of-the-art GNN-based NIDS have fundamentally different vulnerabilities
2. Three attack categories validated across 3 datasets: E-GraphSAGE (99% → 0.6%), GNN-RNIDS (99% → 14.7%)
3. E-GraphSAGE resists simple spoofing but fails with benign injection; GNN-RNIDS vulnerable to almost all attacks
4. Cloud dataset (321 features) provides some resilience but still shows 40-50% f1 score drops
5. Real-world feasible attacks: 5000 spoofed IPs, 50-100 benign flows, 100 link targets - all achievable with standard DDoS techniques

Questions?

References

1. https://www.zayo.com/newsroom/average-ddos-attack-cost-businesses-nearly-half-a-million-dollars-in-2023-according-to-new-zayo-data/
2. https://dl.acm.org/doi/10.1145/3543146.3543171
3. https://dl.acm.org/doi/10.1109/noms54207.2022.9789878
4. https://www.scitepress.org/papers/2018/66398/66398.pdf
5. https://www.sciencedirect.com/science/article/abs/pii/S0167404824004656
6. https://www.yorku.ca/research/bccc/ucs-technical/cybersecurity-datasets-cds/cloud-ddos-attacks-bccc-cpacket-cloud-ddos-2024/

Thank You!