NDSS SYMPOSIUM 2026

San Diego, CA
February 23-27, 2026

CoLD: Collaborative Label Denoising Framework for Network Intrusion Detection

Handling Noisy Labels From Causal Perspective in Network Security

Authors

Shuo Yang†, Xinran Zheng‡, Jinze Li†, Jinfeng Xu†, Edith C. H. Ngai†*

†The University of Hong Kong ‡University College London

*Corresponding Author

shuoyang.ee@gmail.com
chngai@eee.hku.hk

I Introduction: Label Noise in Network Intrusion Detection

The Critical Role of IDS

Intrusion Detection Systems (IDS) are essential for identifying and mitigating malicious activities in network security. Modern IDSs predominantly rely on data-driven models trained on labeled data , where high-quality labels are essential for learning effective representations.

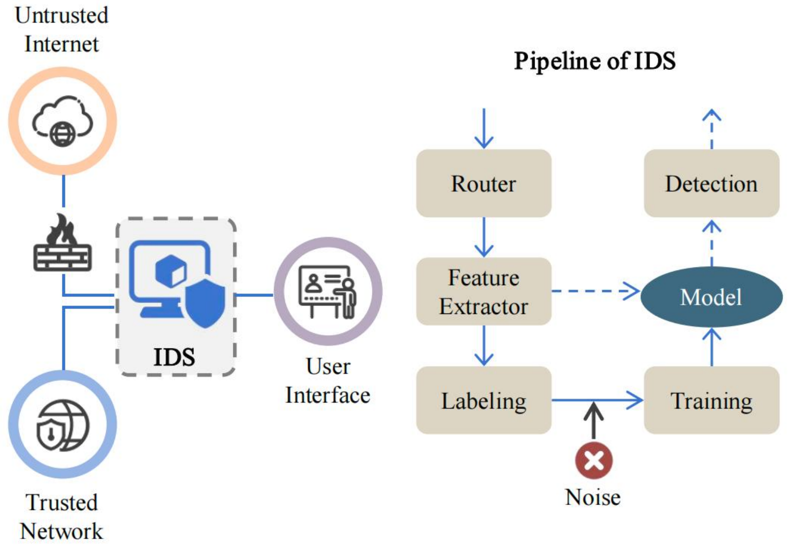
As network traffic grows in complexity and volume, the need for accurate and reliable IDS becomes increasingly urgent.

Severe Impact on Performance

IDS models trained on noisy datasets tend to **perform poorly**, generating false positives for benign actions and failing to detect critical threats. This undermines system reliability and imposes a heavy burden on security teams.

↓ 15-50%
Performance drop at 20-40% noise

High
False positive rate increase



Sources of Label Noise

1. **Human bias and labeling errors** during manual annotation
2. **Dynamic network environments** with evolving attack patterns
3. **Stealth attacks and encryption** mimicking normal patterns
4. **Advanced malware variants** blurring benign/malicious boundaries

Research Question: How can we fundamentally understand and effectively address label noise in network intrusion detection to build more reliable and robust IDS?

2 Problem Statement: Key Challenges

C1 Mechanism Understanding Gap

The Fundamental Issue

While it is well-established that noisy labels negatively impact data-driven models, **the underlying mechanism** of how noisy labels affect learning in network traffic remains poorly understood .

Consequences

- Hinders development of targeted solutions
- Limits model robustness improvements
- Prevents informed architectural design

C2 Existing Method Limitations

Two Main Categories:

Robust Training Methods

Modify loss functions or training strategies to make models resilient to noise.

Limitations: Rely on unrealistic assumptions—prior knowledge of label reliability, access to clean validation data, or known noise transition matrices.

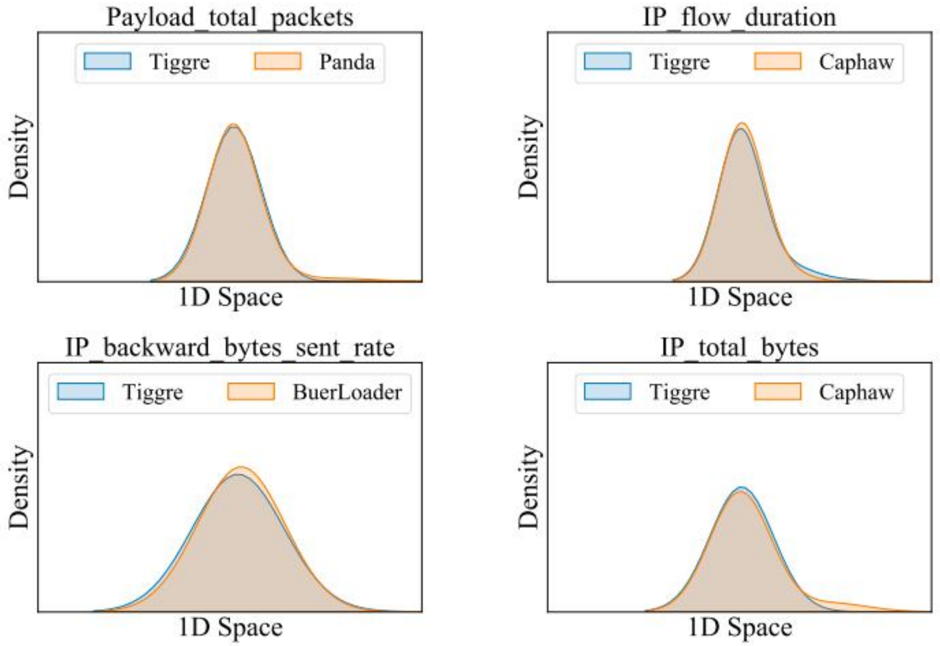
Dataset Purification Methods

Directly detect and correct mislabeled instances using metric learning or active learning.

Limitations: Distance-based measurements struggle with local consistency where features from different categories share similar distributions.

The Local Consistency Problem

Local consistency refers to the phenomenon where features from different categories share similar distributions in the feature space. This occurs because different types of network traffic are often generated under similar conditions.



Our Approach

We address both challenges through:

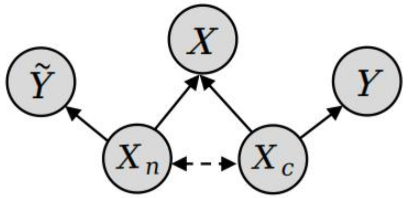
For C1: Causal analysis revealing local consistency promotes spurious associations

For C2: Causal Collaborative Denoising with multi-view representation learning

3 Causal Analysis: Understanding the Root Cause

Structural Causal Model (SCM)

We employ SCM to delineate interactions between features and labels, constructing a detailed causal graph with five key variables:



Xc → Causal Features
Directly determine ground truth label Y

Xn → Non-Causal Features
Influence noisy label Ŷ but not Y

Y → Ground Truth
True label determined solely by Xc

Ŷ → Noisy Label
Observed label influenced by Xn

Causal Pathways & Spurious Associations

True Causal Path

Xc → Y (causal features directly influence ground truth)

Backdoor Path (Bias Source)

Xn ↔ Xc → Y creates spurious associations between non-causal features and ground truth

Xc becomes a confounder, opening a backdoor path that introduces bias

Shortcut with Noisy Labels

Xc ↔ Xn → Ŷ (model learns wrong associations)

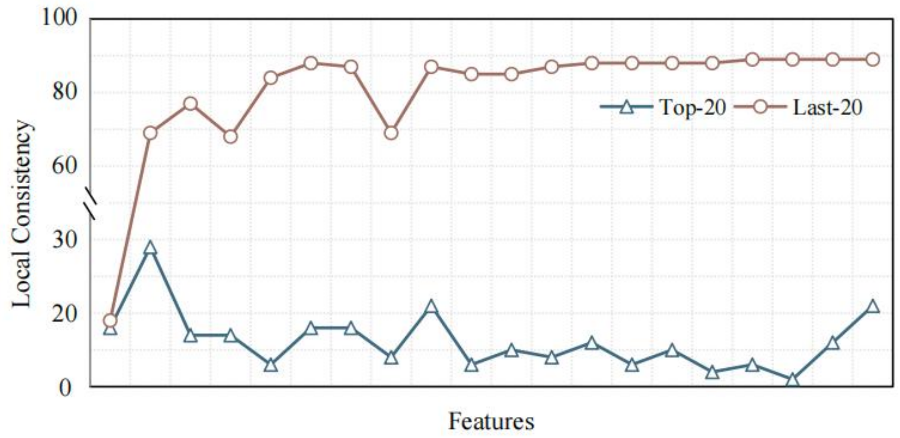
Noisy label Ŷ directs model to focus on Xn, distorting the causal pathway and ignoring Xc → Y

Key Insight

Local consistency amplifies the noise problem by encouraging models to learn **naive, non-discriminative patterns** that fail to capture true decision boundaries.

Empirical Evidence: Local Consistency

Analysis of MALTLS-22 dataset using Kolmogorov-Smirnov test reveals substantial local consistency:

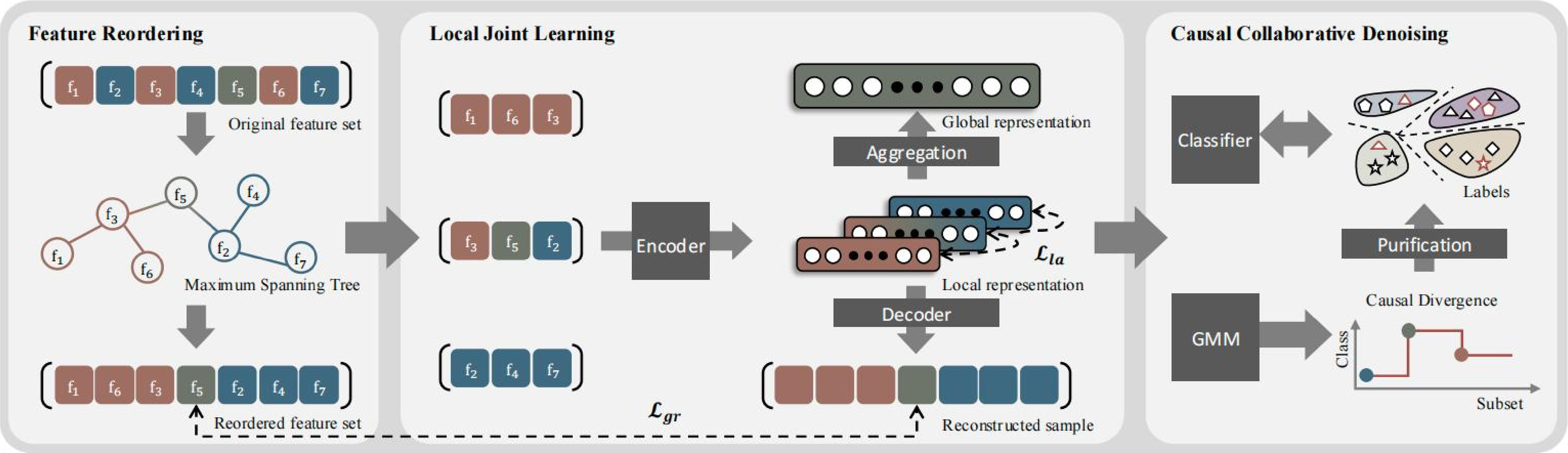


Category combinations of last-20 features share similar distributions(~80+). Even most important features(top-20) exhibit **significant overlap** across categories.

Solution Direction: Disrupt local consistency and suppress spurious associations to improve robustness in noisy environments.

Methodology: Overview of CoLD

CoLD (Collaborative Label Denoising) is a three-component framework designed to enhance robustness of data-driven IDS models in noisy environments by analyzing causal divergences between multiple representations and their potential true labels .



1 Feature Reordering

Objective

Optimize semantic coherence and prepare meaningful subsets for learning

2 Local Joint Learning

Objective

Disrupt local consistency and extract fine-grained, robust, label-independent representations

3 Causal Collaborative Denoising

Objective

Identify and isolate noisy labels by analyzing causal associations

4.1 Methodology: Feature Reordering & Local Joint Learning

Feature Reordering

Partitioning features into subsets requires maximizing local correlations to retain essential semantic information.

Procedure

1. **Compute Correlation Matrix**
   $FCM_{ij} = \frac{\mathrm{Cov}(\mathbf{x}(f_i), \mathbf{x}(f_j))}{\sigma_{f_i} \cdot \sigma_{f_j}}, \forall i, j \in \{1, 2, ..., d\}$
2. **Construct MST**
   Maximum Spanning Tree maximizes total correlation weight
3. **DFS Traversal**
   Determine feature ordering from highest correlation pair

Feature Obfuscation

Apply random masking to increase local feature diversity and disrupt local consistency:

$$\tilde{\mathbf{x}}_i = \mathbf{m} \odot \mathbf{x}_i + (1 - \mathbf{m}) \odot \mathbf{x}_j$$

where m ~ Bernoulli(δ)

Mask vector m randomly samples features, creating perturbed versions for robust learning.

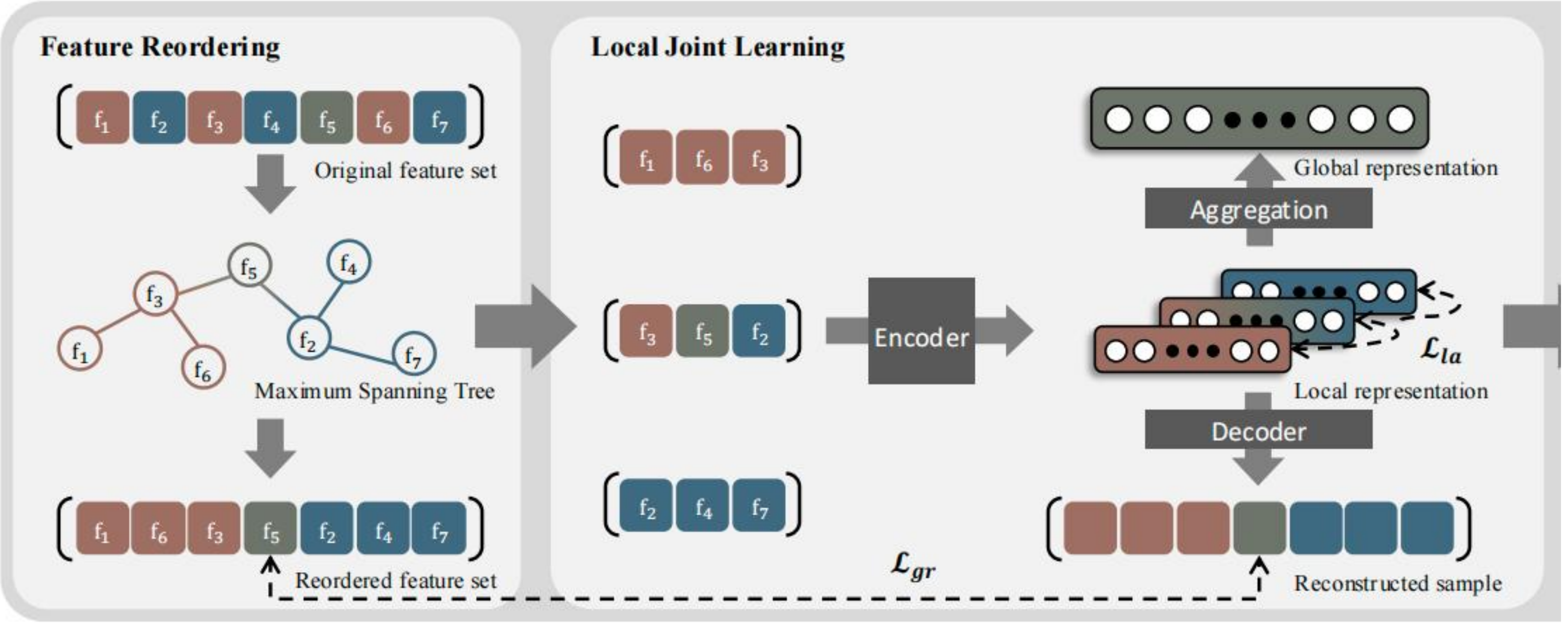
Local Alignment $\mathcal{L}_{la}$

Align representations from different subsets of the same sample using **contrastive learning**:

Global Reconstruction $\mathcal{L}_{gr}$

Ensure local representations reflect overall global structure

Minimizes distance between reconstructed local features and original global features.



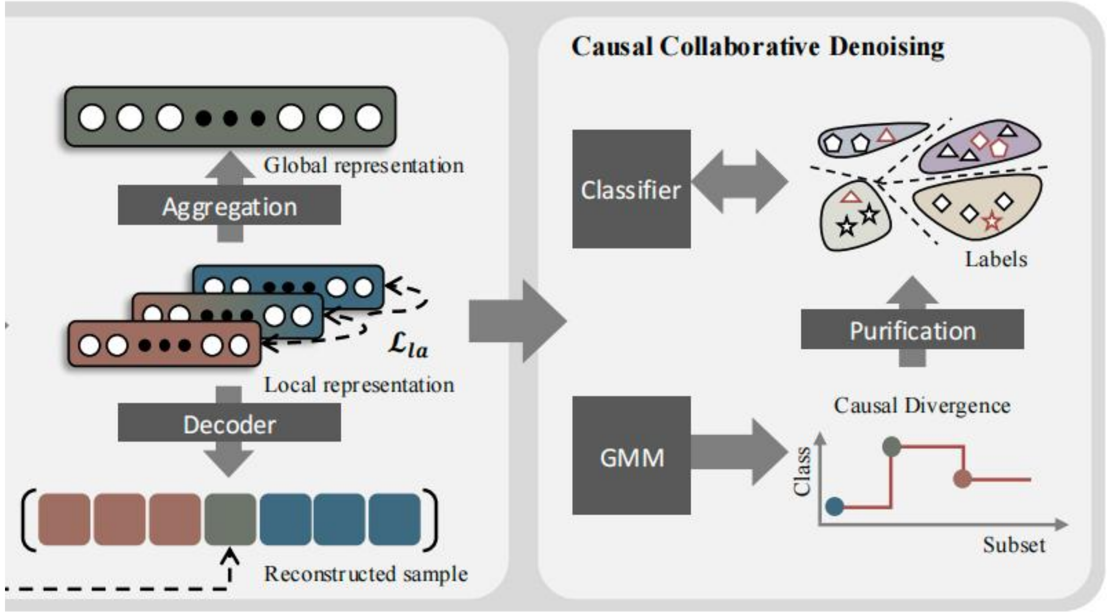
Overall Objective

$$\mathcal{L} = \mathcal{L}_{la} + \mathcal{L}_{gr}$$

Encourages learning discriminative features from multiple perspectives while ensuring global perception.

Self-Supervised Multi-view Learning

Methodology: Causal Collaborative Denoising



Gaussian Mixture Model (GMM)

GMM models complex distributions of network traffic samples with class overlap:

Modeling Process

- Map representation
- Latent variables y ∈ [1, 2, ..., K] assign to mixture components
- Compute posterior probabilities $\gamma_{i,j,k}$

$$\gamma_{i,j,k} = \frac{\pi_k \mathcal{N}\left(\tilde{\mathbf{z}}_{i,j} \mid \mu_k, \sigma_k\right)}{\sum_{l=1}^{K} \pi_l \mathcal{N}\left(\tilde{\mathbf{z}}_{i,j} \mid \mu_l, \sigma_l\right)}$$

Cluster Label Assignment

$\widetilde{y}_{i,j} = \arg\max_k \gamma_{i,j,k}$ for each subset, creating multi-labels per sample

Bridging Gap with Classifier

Link observed labels y_i with GMM predictions through classifier:

$$\theta_h^* = \min_{\theta_h} \left[-\sum_{i=1}^{N} \overline{y}_i \log y_i \right]$$

Update linear head parameters via cross-entropy loss to connect supervised and unsupervised components.

Causal Divergence Metric (CDM)

Quantify probability of noise transfer between multi-labels and observed label:

$$\mathrm{CDM}(\mathbf{x}_i) = \frac{1}{M} \sum_{j=1}^{M} \mathbf{1}(\widetilde{y}_{i,j} \neq y_i | \mathbf{x}_i)$$

where $1(\cdot)$ is indicator function returning 1 if condition is true, 0 otherwise.

Dataset Purification

Threshold-Based Filtering

$$\mathcal{D}_{\mathrm{p}} \leftarrow \mathcal{D} \setminus \{\mathbf{x}_i : \mathrm{CDM}(\mathbf{x}_i) > \epsilon\}$$

We adopt rigorous evaluation with $\epsilon = 0$, ensuring sample retention only if all subsets are causally associated with observed label .

Downstream Training

Purified dataset D_p used to train classifier, ensuring model learns from accurate and representative samples.

5 Experimental Setup

Datasets

MALTLS-22 & CICIDS-2017

Dataset	CICIDS-2017	MALTLS-22
Benign	32.50%	36.94%
Mal. (Head-3)	44.90%	12.33%
Mal. (Tail-3)	9.70%	4.10%
Mal. (Others)	12.9%	46.63%
# of Classes	9	23
Gini coefficient	0.82	0.84

Mal. is the abbreviation of Malicious.

Baseline Methods (8 Total)

Intrusion Detection

ACID: supervised adaptive clustering

CLEID: Comparative learning enhances IDS

Robust Training

Decoupling: Decoupling update strategy

Co-Teaching: Dual network collaborative training

Co-Teaching+: Enhance divergent choices

Dataset Purification

FINE: Feature decomposition and denoising

MORSE: Semi-supervised noise learning

MCRe: Multi-dimensional constraint representation

Noise Settings

Symmetric Noise

Label corruption applied uniformly across both benign and malicious samples

Asymmetric Noise

Corruption exclusively within malicious class, simulating adversarial scenarios where attacks disguise as benign

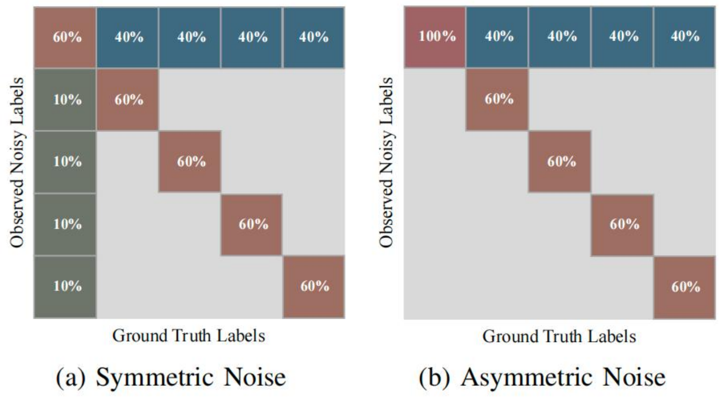


Fig. 6: Example of Label Conversion Matrices with Noise Ratio of 40%.

Main Results: Performance Analysis

CoLD's Superior Robustness

CoLD demonstrates **superior and consistent performance** across all settings, with robustness becoming increasingly evident as noise levels rise.

MALTLS-22 @ 60% Sym: Performance drop only **3.4%** vs 20% noise

CICIDS-2017 @ 60% Sym: Performance drop only **1.6%** vs 20% noise

Intrusion Detection Methods

ACID and CLEID exhibit **subpar performance** with rapid deterioration under noise:

ACID: 15.75% drop (20% → 40% Sym on MALTLS-22)

CLEID: 53.68% drop (20% → 40% Sym on MALTLS-22)

Robust Training Methods

Struggle significantly in **high-noise environments**:

Co-Teaching+: 90%+ → <10% at 60% noise

Lack domain knowledge extraction and positive feedback mechanisms in high-noise scenarios

Dataset Purification Methods

FINE, MORSE, MCRe show greater resilience than robust training methods, but still fall short of CoLD.

MCRe achieves notable results but loses clean samples due to distance-based detection.

TABLE III: Results on MALTLS-22 Dataset.

Noise Type	None	Symmetric					Asymmetric				
Noise Ratio	0%	10%	20%	40%	50%	60%	10%	20%	40%	50%	60%
ACID	92.43/1.83	81.51/3.21	77.46/4.55	61.71/3.84	31.74/4.02	4.38/1.32	80.01/1.85	79.63/2.18	69.48/3.65	39.30/1.78	2.46/0.19
CLEID	91.42/1.12	80.98/0.02	60.61/1.02	6.93/1.74	2.73/0.14	2.38/0.03	85.39/0.05	60.71/1.25	4.93/0.54	3.42/0.35	2.35/0.03
Decoupling	91.30/0.62	89.12/0.79	88.11/1.13	72.66/2.02	31.73/2.35	3.01/1.14	89.54/0.43	90.00/0.37	75.18/0.90	38.37/4.04	3.54/0.42
Co-Teaching	93.47/0.16	92.85/0.17	87.26/0.19	47.50/0.80	7.95/0.53	2.85/0.15	92.57/0.07	90.75/0.29	73.25/0.64	32.98/4.74	2.66/0.11
Co-Teaching+	91.12/0.31	89.49/0.26	89.18/0.56	74.73/0.60	35.50/1.64	3.50/0.12	90.33/0.15	89.23/0.53	86.59/1.71	44.86/5.47	3.02/0.06
FINE	75.76/0.13	64.97/1.35	65.61/0.54	65.37/0.46	57.21/0.27	46.91/1.54	65.43/0.01	64.96/0.51	61.69/1.00	59.19/1.54	59.04/1.04
MORSE	82.04/1.46	77.91/0.13	76.33/0.71	75.71/0.13	74.39/0.85	74.71/1.20	79.36/0.09	77.63/1.90	74.13/0.28	73.92/2.77	70.13/1.09
MCRe	88.49/3.18	87.73/1.94	88.19/0.68	87.03/0.44	86.96/0.38	86.07/0.82	85.66/1.39	85.56/0.73	85.49/0.64	84.97/0.83	84.37/1.13
CoLD (Ours)	92.97/0.32 p=0.043	93.11/0.19 p=0.017	92.14/0.53 p=0.000	91.82/0.42 p=0.000	90.07/0.67 p=0.000	88.75/0.76 p=0.000	93.55/0.34 p= 0.002	91.91/0.35 p=0.000	90.84/0.38 p=0.000	88.08/0.40 p=0.003	86.48/0.65 p=0.008

Real-World Evaluation

Enterprise Network Evaluation

Application scenarios:
Advanced persistent threat (APT) detection in enterprise networks, traceability based IDS system

Integration method:
Using decoupling method, CoLD serves as a plug-in module to enhance the training of existing IDS classifiers

Dataset:
OpTC dataset: Large-scale enterprise network logs containing real-life attack scenarios

Baseline IDS

Flash:
Flash employed Word2Vec to transform node attributes into semantically rich, time-sensitive feature vectors and then utilizes Graph Neural Networks to capture both local and global graph structures. This enables the model to effectively encode complex temporal dependencies within the provenance graph.

Argus:
Argus introduced a dynamic graph representation learning framework that integrates Graph Convolutional Networks with Long Short-Term Memory networks for feature extraction. By embedding timestamp information and supporting dynamic updates, Argus can track and model real-time changes in graph topology.

Classifier
XGBoost (lightweight and efficient to meet real-time processing requirements)

Evaluation Results

TABLE VIII: Results on OpTC Dataset.

Method	Sym-10%	Sym-40%	Asym-10%	Asym-40%
Flash	93.57	79.49	94.08	85.15
Flash+CoLD	94.04 ↑ 0.47	84.89 ↑ 5.40	94.30 ↑ 0.22	93.78 ↑ 8.63
Argus	91.45	81.81	93.94	86.28
Argus+CoLD	93.73 ↑ 2.28	87.83 ↑ 6.02	94.70 ↑ 0.76	93.51 ↑ 7.23

Flash + CoLD @ Sym-40%
84.89% vs 79.49%
↑ 5.40% improvement

Argus + CoLD @ Sym-40%
87.83% vs 81.81%
↑ 6.02% improvement

CoLD consistently improves performance across all noise settings

8

Conclusion & Future Work

Summary

CoLD addresses label noise challenges through **causal analysis**, identifying local consistency as the root cause of performance degradation in network intrusion detection.

Key Contributions

1. **First causal analysis** of noisy labels in network traffic, revealing how local consistency promotes spurious associations
2. **Novel collaborative denoising framework** integrating self-supervised learning with causal inference
3. **Superior performance** across benchmark datasets, significantly outperforming 8 state-of-the-art baselines
4. **Successful real-world deployment** in enterprise networks with 5-6% performance improvements

Future Directions

Advanced Feature Reordering

Explore nonlinear and higher-order dependency capture beyond Pearson correlation

Efficiency Improvements

Reduce computational complexity for large-scale deployment

Streaming Data Adaptation

Integrate with incremental/continual learning for online IDS scenarios

Impact

CoLD paves the way for **more reliable and secure network infrastructures** in noisy environments, enabling robust intrusion detection even when high-quality labeled data is unavailable.

> "By understanding and addressing the root causes of label noise, we can build intrusion detection systems that are truly resilient to the challenges of real-world network environments."

THANKS

CoLD: Collaborative Label Denoising Framework for Network Intrusion Detection

For any questions, please contact:
shuoyang.ee@gmail.com