Cleaning Huge Anomaly-Polluted Log Data Sets Using Sample Selection

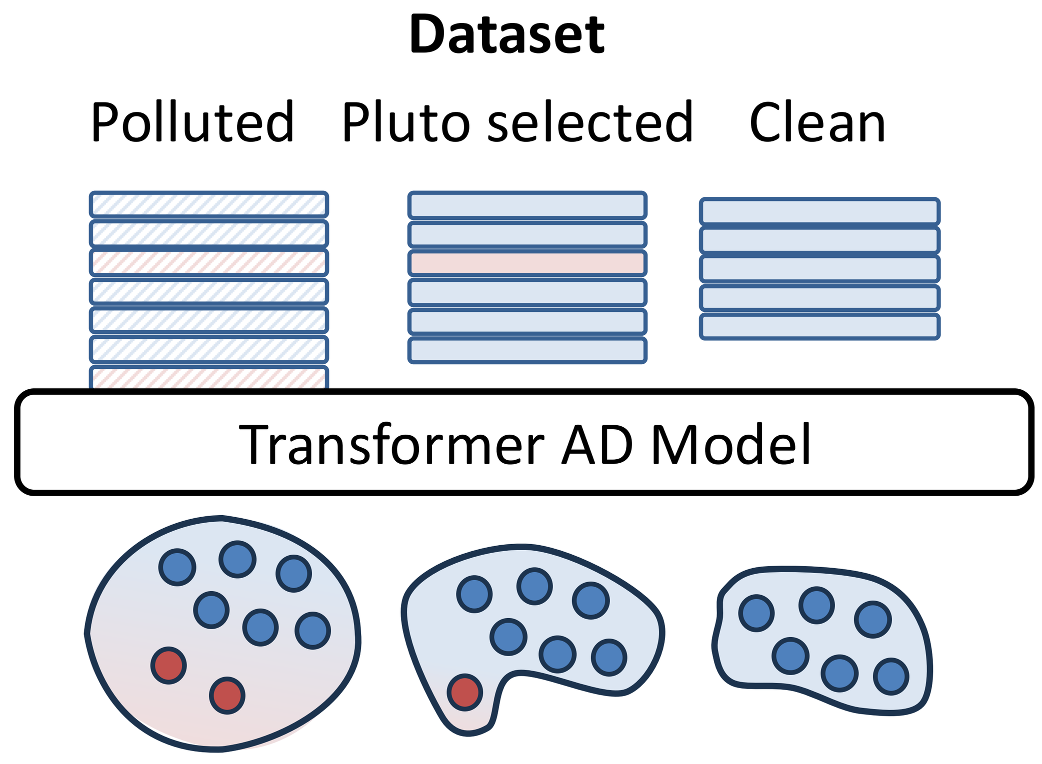
Lei Ma, Lei Cao, Peter M. Vannostrand, Dennis M. Hofmann, Yao Su, and Elke A. Rundensteiner

1 MOTIVATION

State-of-the-art log anomaly detection methods typically depend on a clean dataset of log sequences containing only normal data, which requires costly human labeling efforts. In contrast, using a polluted dataset (unlabeled data with anomalies) can severely degrade model performance due to overfitting to anomalies.

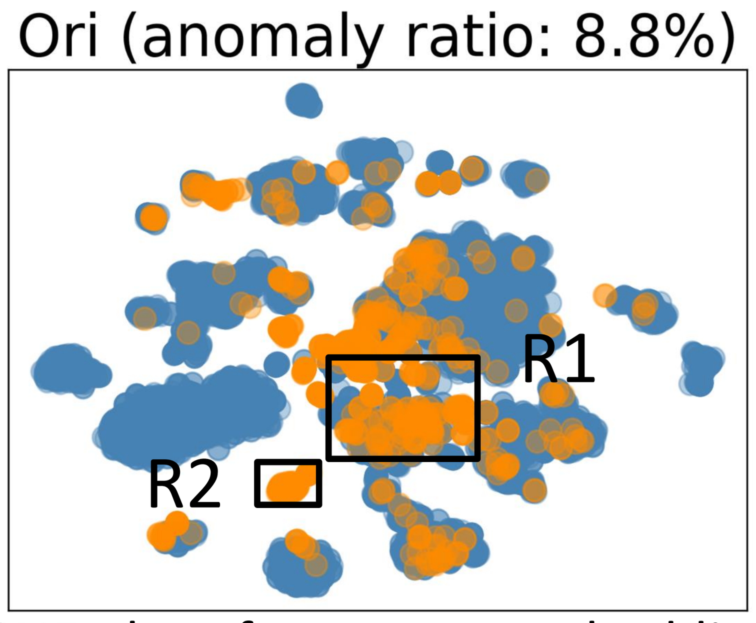
This work focuses on leveraging the characteristics of the embedding space to identify and select a clean subset of normal sequences from polluted data, which is then used to train a Transformer-based anomaly detection model.

This talk is Based on the paper "**Pluto: Sample Selection for Robust Anomaly Detection on Polluted Log Data**"[1] accepted at SIGMOD 25.



2 CHALLENGES

Ori (anomaly ratio: 8.8%)



T-SNE plot of sequence embeddings BGL dataset

Uneven Global Pollution

Anomalies are not distributed evenly

Anomaly Subtlety with Slight Pollution

Anomalies can be similar to normal data in slight polluted region (R1: anomaly ratio 20.9%)

Anomaly Concentration with High Pollution

Anomalies can be similar to each in highly polluted region (R2: anomaly ratio 100%)

3 STATE-OF-THE-ART

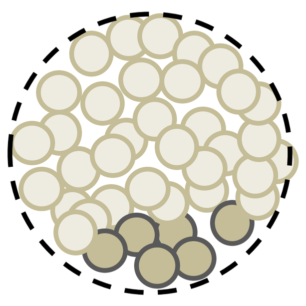
Sample Selection Methods (Co-teaching[2], FINE[3], ITLM[4]) select clean data from a noisy dataset, based on two assumptions:

- Assumption 1: Random Noise distinguished from clean data
- Assumption 2: Evenly distributed Noise

None of our challenges satisfies these assumptions.

5 METHODOLOGY

Local Pollution Level Estimation



Embedding Cluster

Empirical SVD

$E \approx \lambda_1 \boldsymbol{u_1} \cdot \boldsymbol{v_1} + \lambda_2 \boldsymbol{u_2} \cdot \boldsymbol{v_2}$ (1)

dom = $\frac{\lambda_1}{\lambda_2}$

Unknown Ground Truth

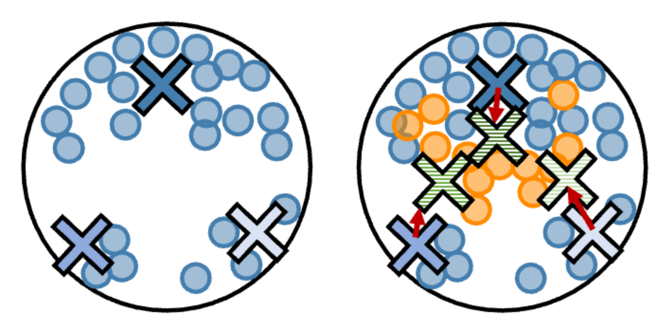
$E \approx \lambda_- \boldsymbol{u_-} \cdot \boldsymbol{v_-} + \lambda_+ \boldsymbol{u_+} \cdot \boldsymbol{v_+}$ (2)

Estimate?

Pollution = $\frac{\lambda_-}{\lambda_+}$

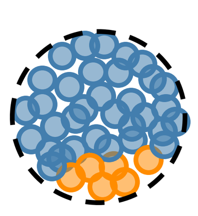
Spectrum-purifying Selection Strategy

Anomaly Perturbance to Eigenvectors



Without anomalies　With anomalies

Eigenvectors are perturbed, **differently**.



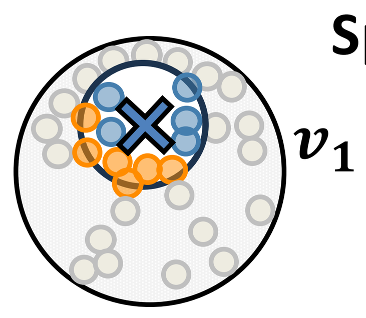
Case 1: Slightly Polluted

First component ↔ Normal

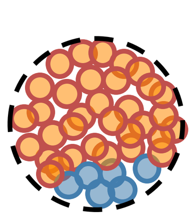
Second component ↔ Abnormal

Pollution $\sim \frac{1}{dom}$ Low dominance

Spectrum-picking Strategy



Selecting samples close to the first eigenvector $\boldsymbol{v_1}$



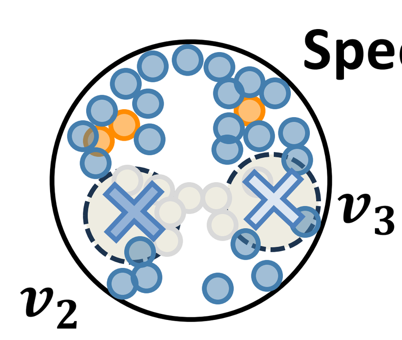
Case 2: Highly Polluted

First component ↔ Abnormal

Second component ↔ Normal

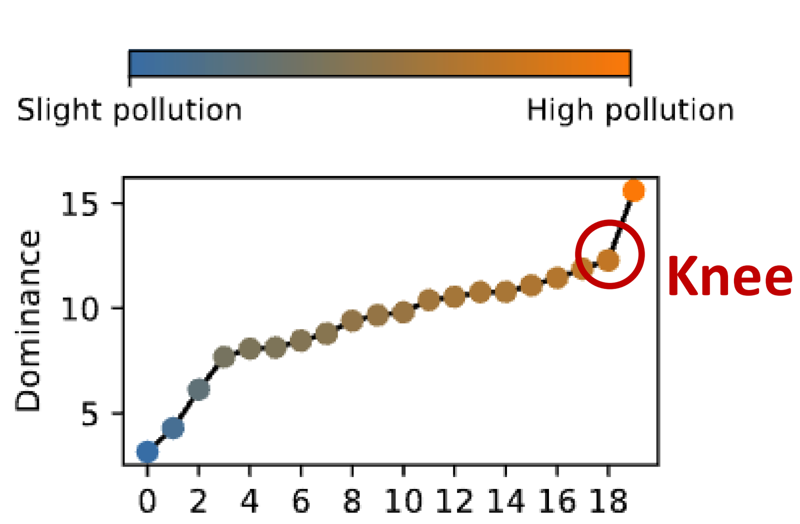
Pollution $\sim dom$ High dominance

Spectrum-purifying Strategy



Discarding samples close to the minor eigenvector $\boldsymbol{v_{2,3}}$...

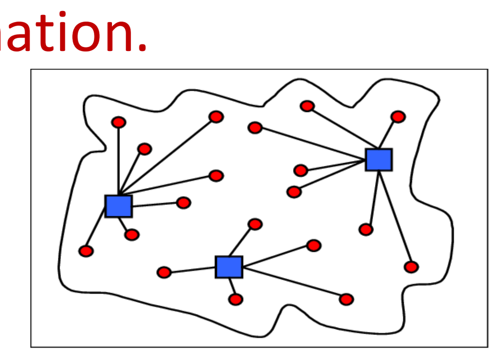
High Pollution Cluster Detection



Dominance of all clusters

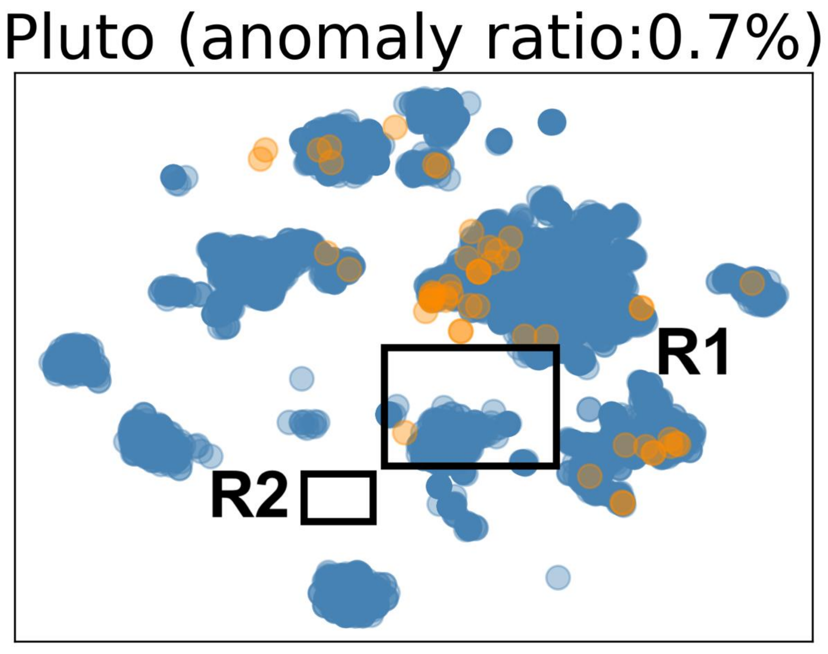
K-Medoid, Facility location

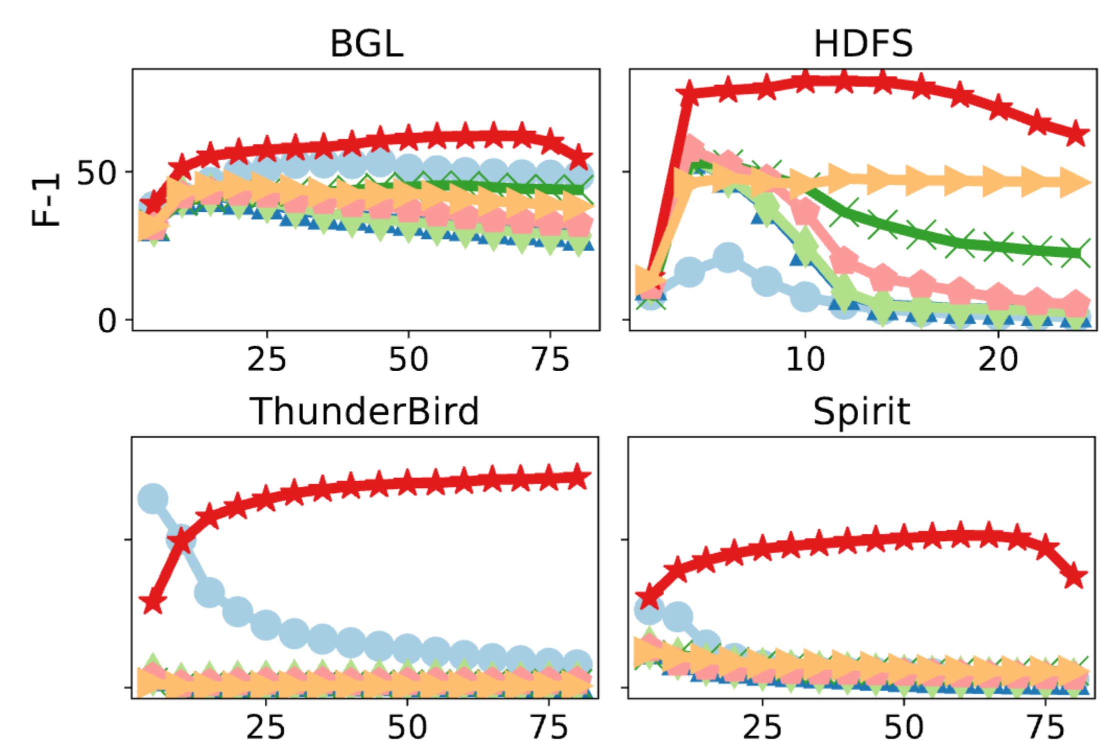
Can be transformed to a k-medoid problem[5], solved by greedy with $1 - \frac{1}{e}$ approximation.



6 EXPERIMENTS

Pluto (anomaly ratio:0.7%)





Top t log keys

7 REFERENCE

[1] Lei Ma, Lei Cao, Peter M. VanNostrand, Dennis M. Hofmann, Yao Su, and Elke A. Rundensteiner. 2024. Pluto: Sample Selection for Robust Anomaly Detection on Polluted Log Data. Proc. ACM Manag. Data 2, 4, Article 203 (Sept. 2024)

[2] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. 2018. Co-teaching: Robust training of deep neural networks with extremely noisy labels. NeulPS 31 (2018)

[3] Taehyeon Kim, Jongwoo Ko, JinHwan Choi, Se-Young Yun, et al. 2021. Fine samples for learning with noisy labels. Advances in NeulPS 34 (2021), 24137–24149.

[4] Yanyao Shen and Sujay Sanghavi. 2019. Learning with bad training data via iterative trimmed loss minimization. In ICML. PMLR

[5] Baharan Mirzasoleiman, Kaidi Cao, and Jure Leskovec. 2020. Coresets for robust training of deep neural networks against noisy labels. Advances in Neural Information Processing Systems 33 (2020), 11465–11477.