Cleaning Huge Anomaly-Polluted Log Data Sets Using Sample Selection

Lei Ma

PhD Candidate, WPI

About This Work

This talk is based on the SIGMOD 25 paper:

Pluto: Sample Selection for Robust Anomaly Detection on Polluted Log Data.

LEI MA[1], LEI CAO[2], PETER M. VANNOSTRAND[1], DENNIS M. HOFMANN[1], YAO SU[1], and ELKE A. RUNDENSTEINER[1]

1 Worcester Polytechnic Institute

2 University of Arizona

Outline

- **Background**
- Motivation
- Pluto Overview
  - Local Pollution Estimation
  - Selection Strategy
- Experimental Evaluation
- Conclusion

Log and Log Sequences

Index	Log Type	Log Message
1	A	connection from <*ip*>
2	B	error: cannot connect to <*ip*>
3	C	session closed

Log

Log and Log Sequences

Index	Log Type	Log Message
1	A	connection from *<ip>*
2	B	error: cannot connect to *<ip>*
3	C	session closed

Log

Log Sequence

A	B	C

Symbolic Perspective

Log and Log Sequences

Index	Log Type	**Log Message**
1	A	connection from *<ip>*
2	B	error: cannot connect to *<ip>*
3	C	session closed

Log

Log Sequence

A	**B**	C

Symbolic Perspective

Pluto: This Work

connection from *<ip>.* error: cannot connect to *<ip>.* session closed.

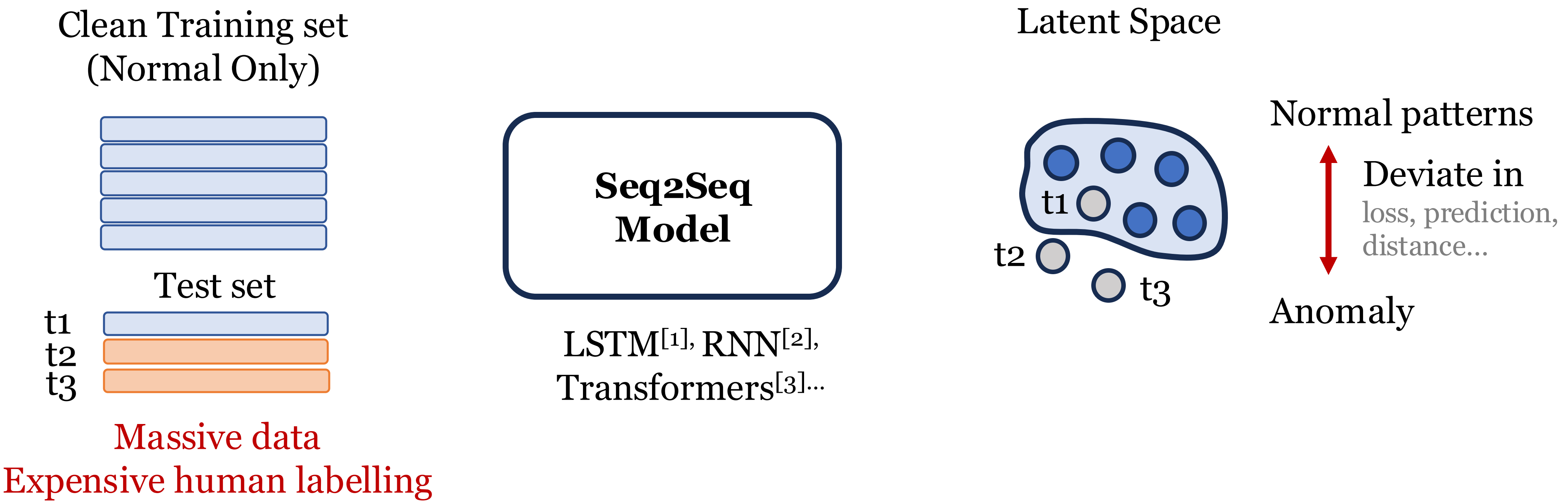
Semantic Perspective

Krone: Ongoing work with LLM

Outline

- Background
- **Motivation**
- Pluto Overview
  - Local Pollution Estimation
  - Selection Strategy
- Experimental Evaluation
- Conclusion

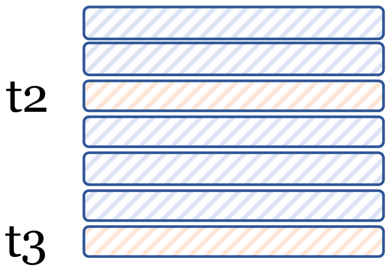
Traditional One-class Log Anomaly Detection



[1] Min Du, Feifei Li, Guineng Zheng, and Vivek Srikumar. 2017. Deeplog: Anomaly detection and diagnosis from system logs through deep learning. In 2017 ACM SIGSAC. 1285–1298.
[2] Zhiwei Wang, Zhengzhang Chen, Jingchao Ni, Hui Liu, Haifeng Chen, and Jiliang Tang. 2021. Multi-scale one-class recurrent neural networks for discrete event sequence anomaly detection. In ACM SIGKDD.
[3] Haixuan Guo, Shuhan Yuan, and Xintao Wu. 2021. Logbert: Log anomaly detection via bert. In 2021 International Joint Conference on Neural Networks (IJCNN). IEEE, 1–8

Traditional One-class Log Anomaly Detection

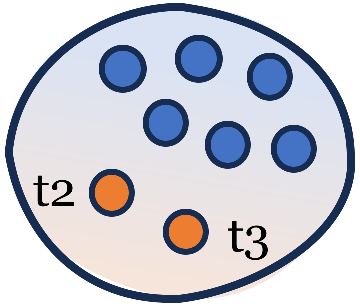
Unlabeled dataset
(Polluted)



Seq2Seq Model

LSTM[1], RNN[2], Transformers[3]...

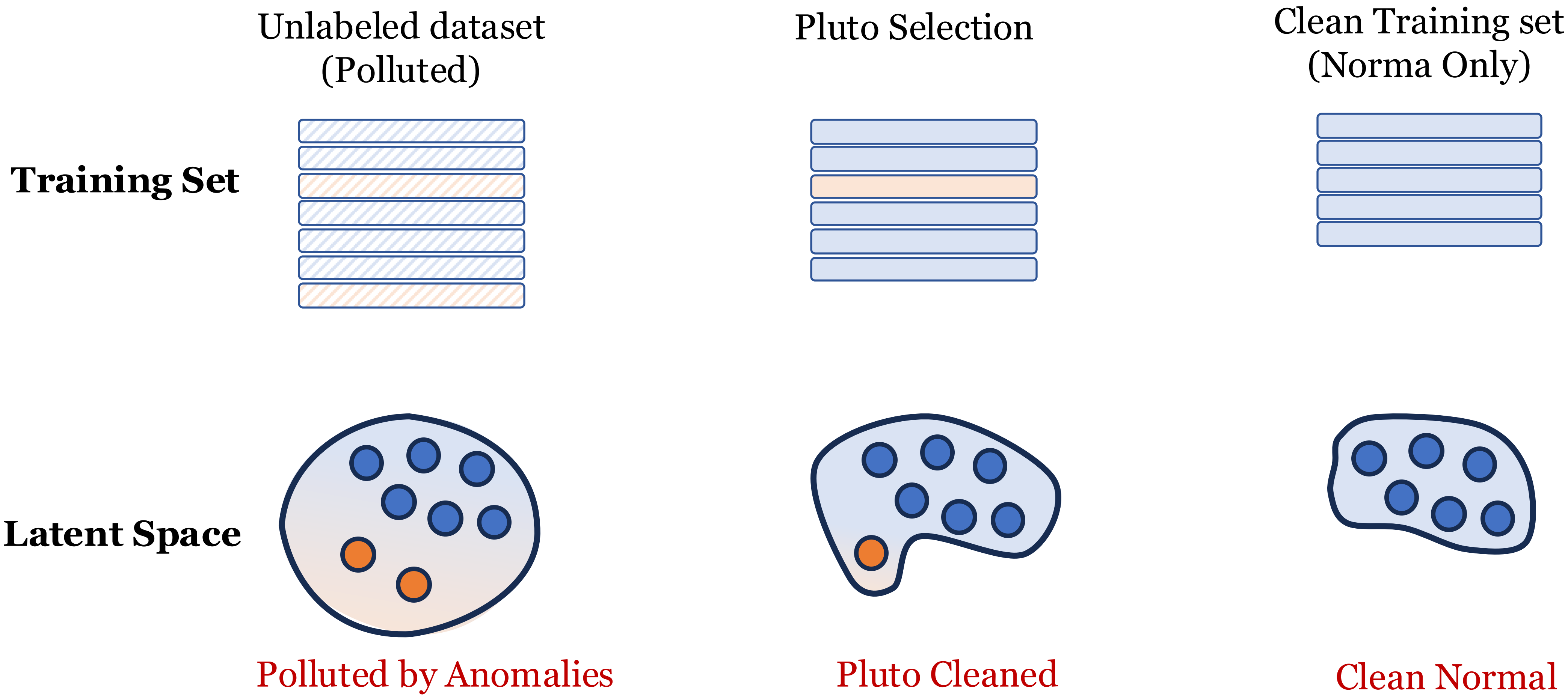
Latent Space



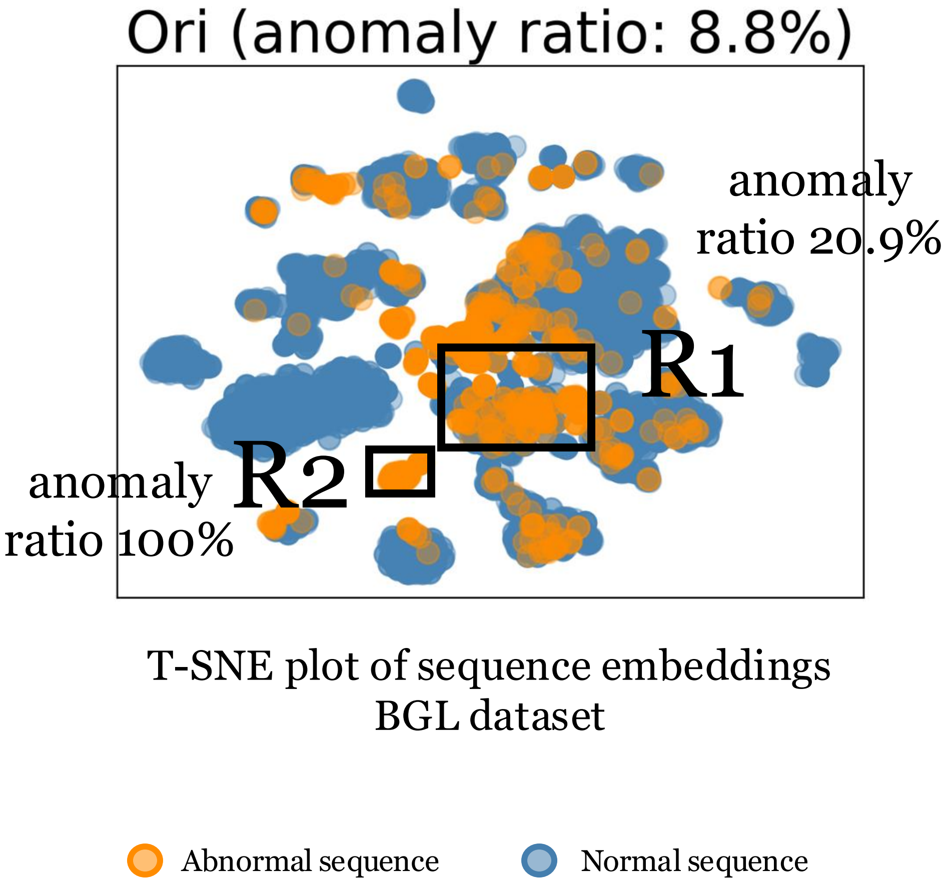
Corrupted learned patterns
Bad detection performance

[1] Min Du, Feifei Li, Guineng Zheng, and Vivek Srikumar. 2017. Deeplog: Anomaly detection and diagnosis from system logs through deep learning. In 2017 ACM SIGSAC. 1285–1298.
[2] Zhiwei Wang, Zhengzhang Chen, Jingchao Ni, Hui Liu, Haifeng Chen, and Jiliang Tang. 2021. Multi-scale one-class recurrent neural networks for discrete event sequence anomaly detection. In ACM SIGKDD.
[3] Haixuan Guo, Shuhan Yuan, and Xintao Wu. 2021. Logbert: Log anomaly detection via bert. In 2021 International Joint Conference on Neural Networks (IJCNN). IEEE, 1–8

Motivation of Sample Selection



Challenges and SOTA



T-SNE plot of sequence embeddings
BGL dataset

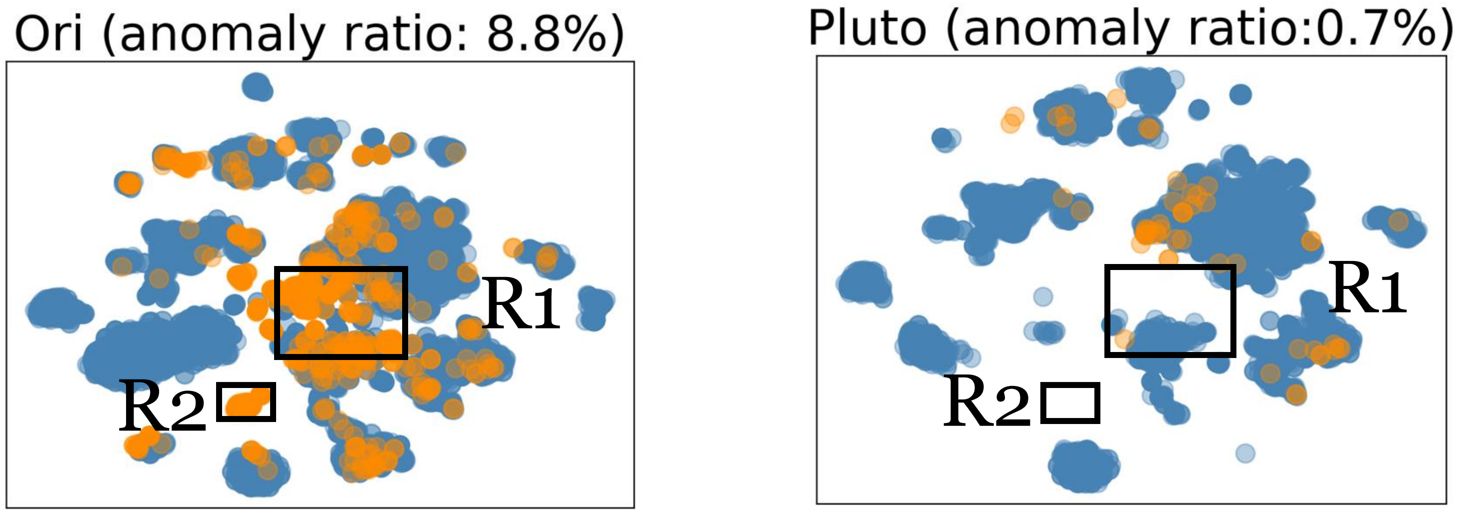
- **CH1: Uneven Global Pollution**
- **CH2: Anomaly Subtlety with Slight Pollution**
  Anomalies can be similar to normal data
- **CH3: Anomaly Concentration with High Pollution**
  Anomalies can be similar to each other

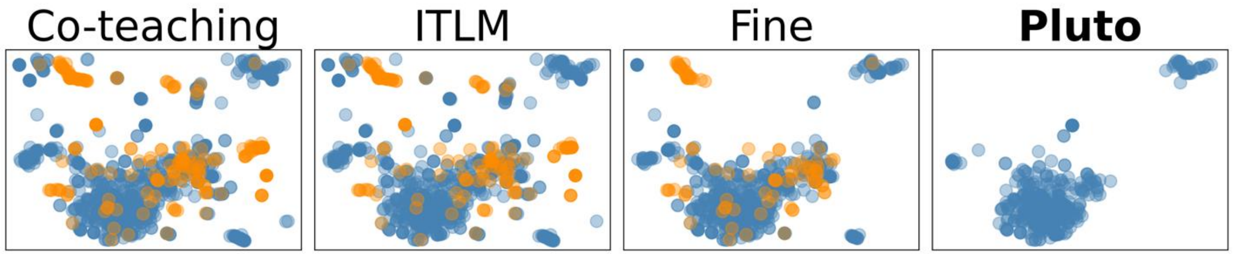
SOTA selection methods (Co-teaching[1],FINE[2],ITLM[3]) select clean data from a noisy dataset

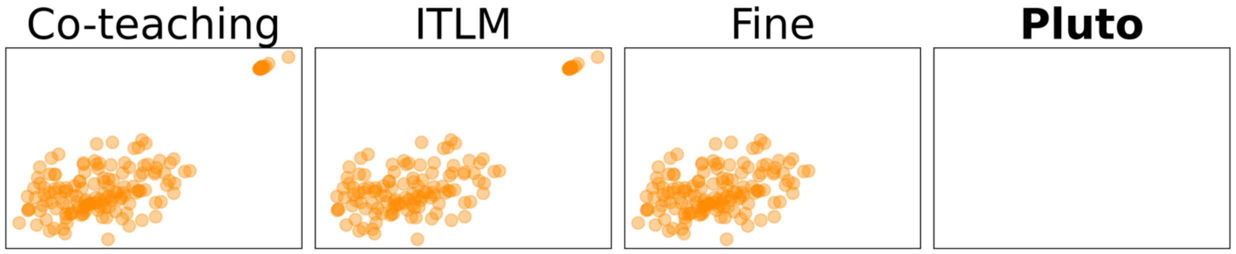
- **Assumption 1: Random Noise distinguished from clean data**
- **Assumption 2: Evenly distributed Noise**

[1] 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)
[2] 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.
[3] Yanyao Shen and Sujay Sanghavi. 2019. Learning with bad training data via iterative trimmed loss minimization. In ICML. PMLR

Pluto Selection Results - Visualization







R1: Anomaly Subtlety Region

SOTA vs. Pluto selection

Anomaly ratio 20.9% (original) → 0.2% (Pluto).

R2: Anomaly Concentration Region

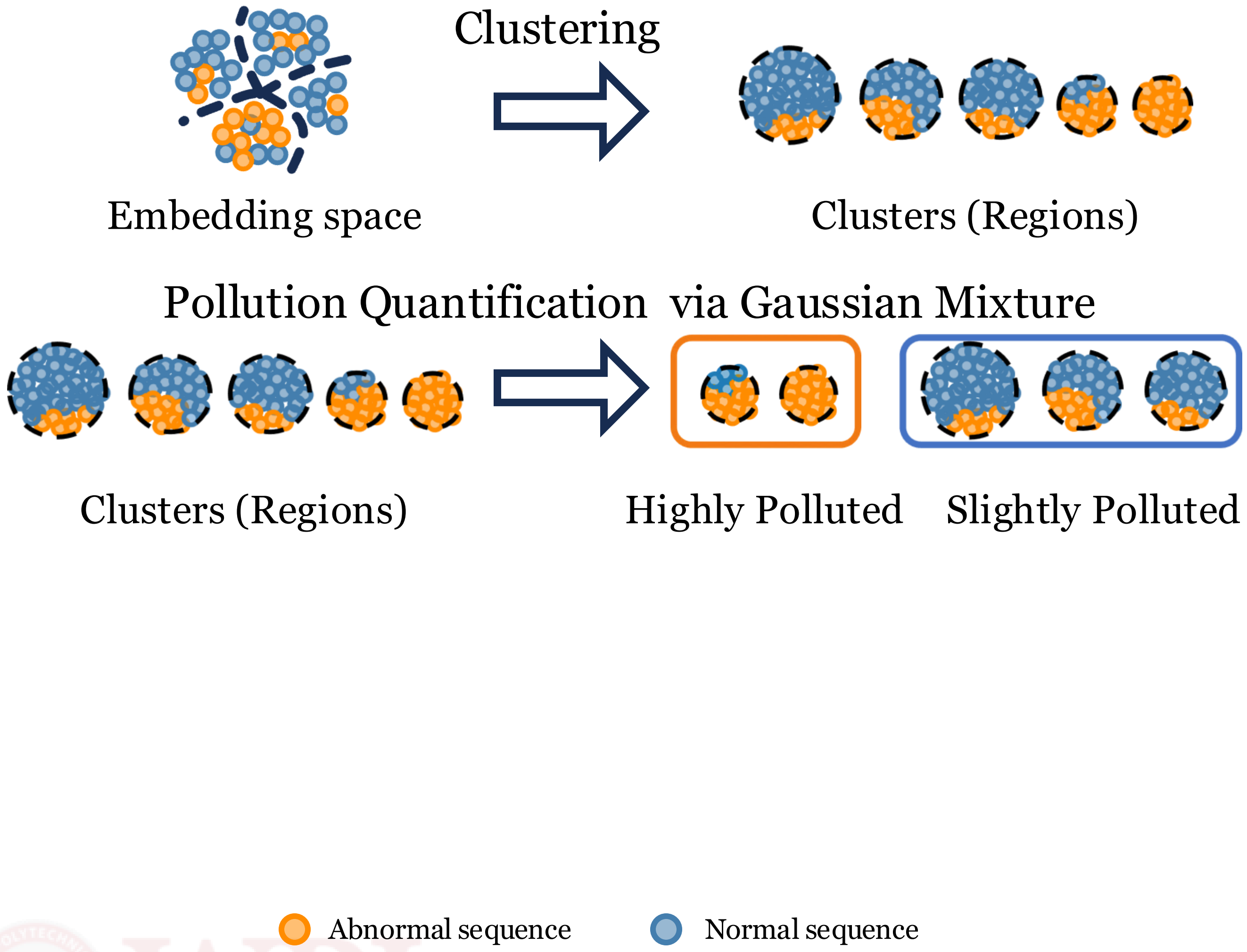
SOTA vs. Pluto selection

Anomaly ratio 100.0% (original) → 0.0% (Pluto)

Outline

- Background
- Motivation
- **Pluto Overview**
  - Local Pollution Level Estimation
  - Selection Strategy
- Experimental Evaluation
- Conclusion

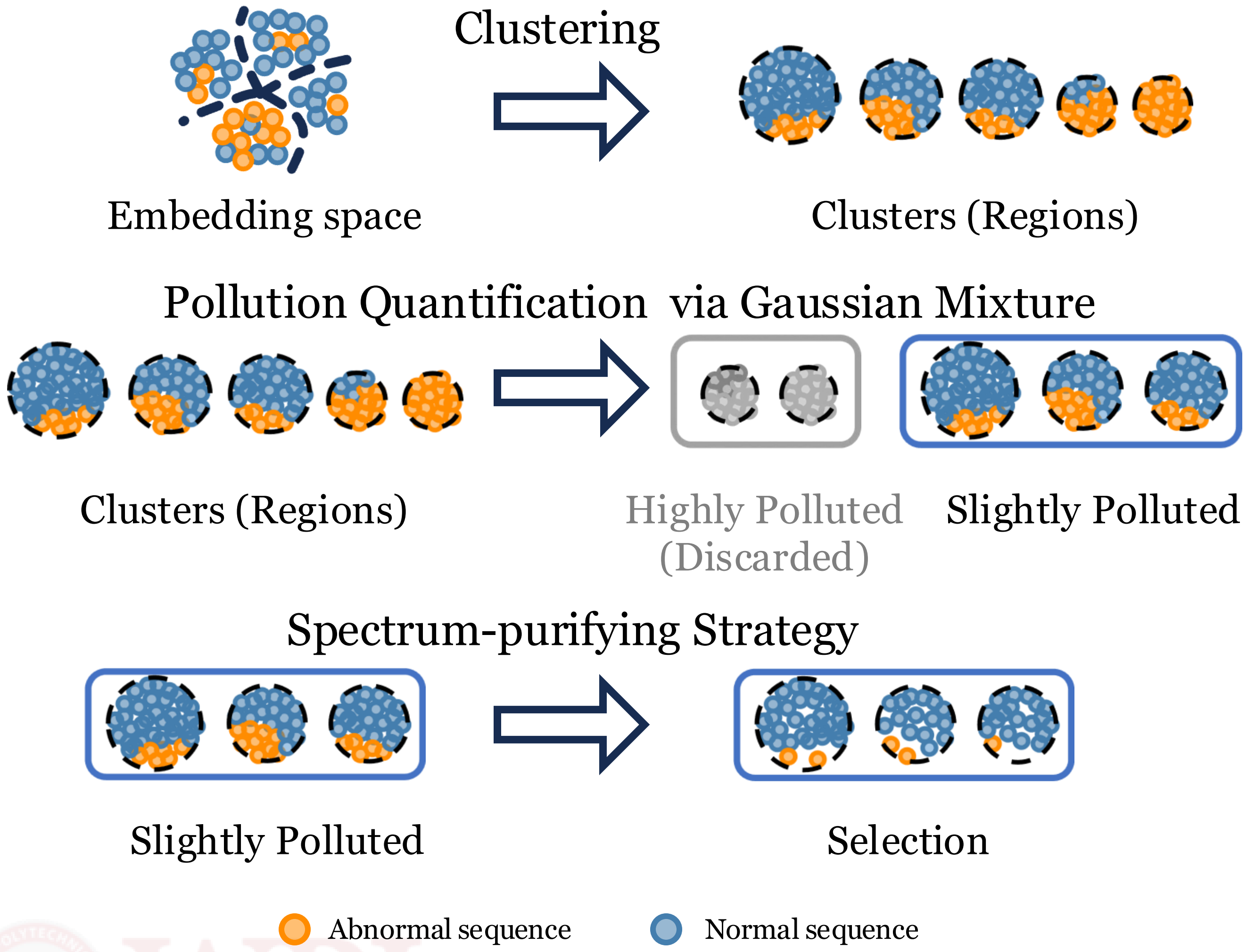
Pluto Overview



- **Region Partitioning**
  CH1: Uneven Global Pollution
  Clustering algorithm

- **Local Pollution Level Estimation**
  CH2: Anomaly Concentration
  Estimate the pollution level of the cluster, discard highly polluted ones

Pluto Overview



- **Region Partitioning**
  CH1: Uneven Global Pollution
  Clustering algorithm

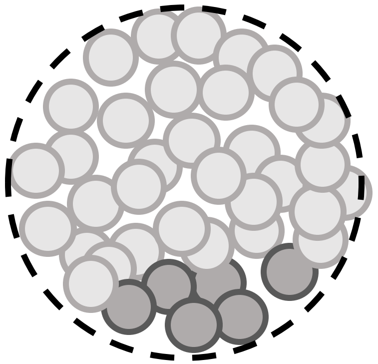
- **Local Pollution Level Estimation**
  CH2: Anomaly Concentration
  Estimate the pollution level of the cluster, discard highly polluted ones

- **Sample Selection Strategy**
  CH3: Anomaly Subtlety
  Sample selection in slightly polluted clusters

Outline

- Background
- Motivation
- Pluto Overview
  - **Local Pollution Level Estimation**
  - Selection Strategy
- Experimental Evaluation
- Conclusion

Pollution Level Estimator



Cluster of embeddings

Empirical SVD

$E \approx \lambda_1 \boldsymbol{u_1} \cdot \boldsymbol{v_1} + \lambda_2 \boldsymbol{u_2} \cdot \boldsymbol{v_2}$ (1) λ_1, λ_2: empirical eigenvalues ($\lambda_1 > \lambda_2$)

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

The empirical dominance of the **first component** to the **second component**

Unknown Ground Truth

$E \approx \lambda_- \boldsymbol{u}_- \cdot \boldsymbol{v}_- + \lambda_+ \boldsymbol{u}_+ \cdot \boldsymbol{v}_+$ (1) λ_-, λ_+: eigenvalues of abnormal and normal components

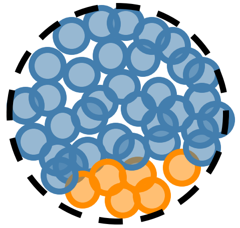
$$\text{Pollution} = \frac{\boldsymbol{\lambda_-}}{\boldsymbol{\lambda_+}}$$

Estimate?

The dominance of the **abnormal component** to the **normal component**

Pollution Level Estimator

$$\text{Pollution} = \frac{\lambda_-}{\lambda_+} \xleftarrow{?} \text{dom} = \frac{\lambda_1}{\lambda_2}$$



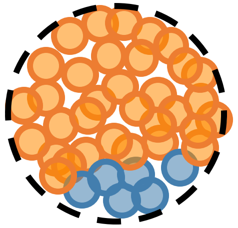
Case 1: Slightly polluted

First component ↔ Normal
Second component ↔ Abnormal

$\lambda_1 \rightarrow \lambda_+$
$\lambda_2 \rightarrow \lambda_-$

$$\text{Pollution} \sim \frac{\mathbf{1}}{\boldsymbol{dom}}$$

Low dominance



Case 2: Highly polluted

First component ↔ Abnormal
Second component ↔ Normal

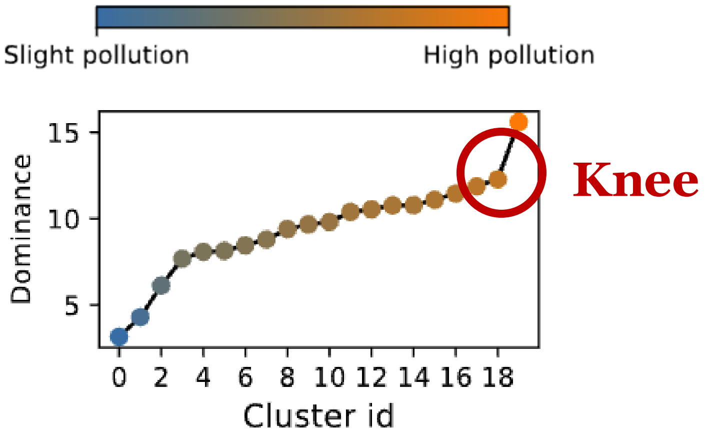
$\lambda_1 \rightarrow \lambda_-$
$\lambda_2 \rightarrow \lambda_+$

$$\text{Pollution} \sim \boldsymbol{dom}$$

High dominance

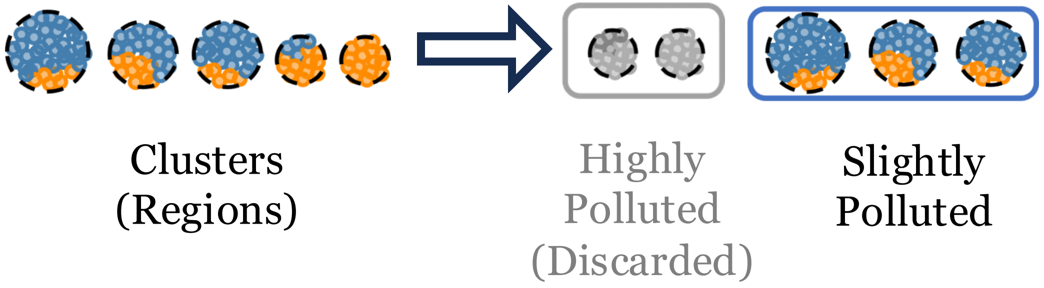
High Pollution Cluster Detection

Step 1: High pollution cluster detection



Dominance of all clusters

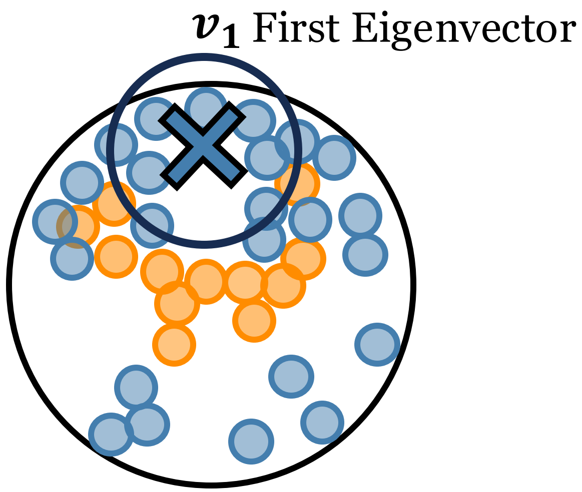
Step2: Discard



Outline

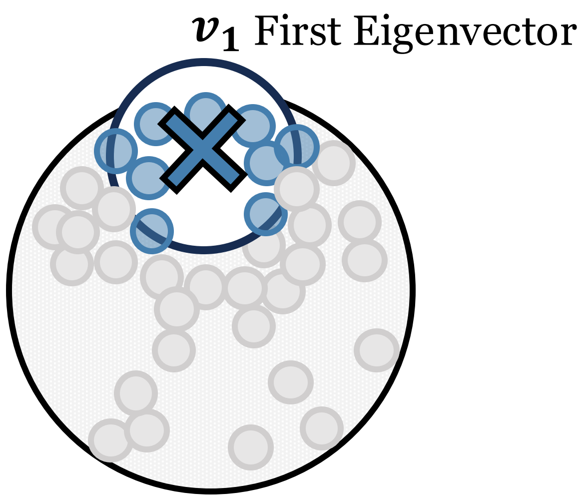
- Background
- Motivation
- Pluto Overview
  - Local Pollution Level Estimation
  - **Selection Strategy**
- Experimental Evaluation
- Conclusion

Ideal Spectrum-picking Strategy



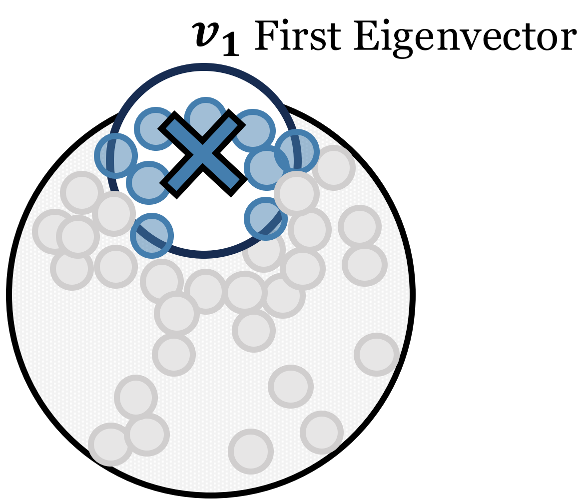
Ideal Spectrum-picking Strategy

Ideal Spectrum-picking Strategy

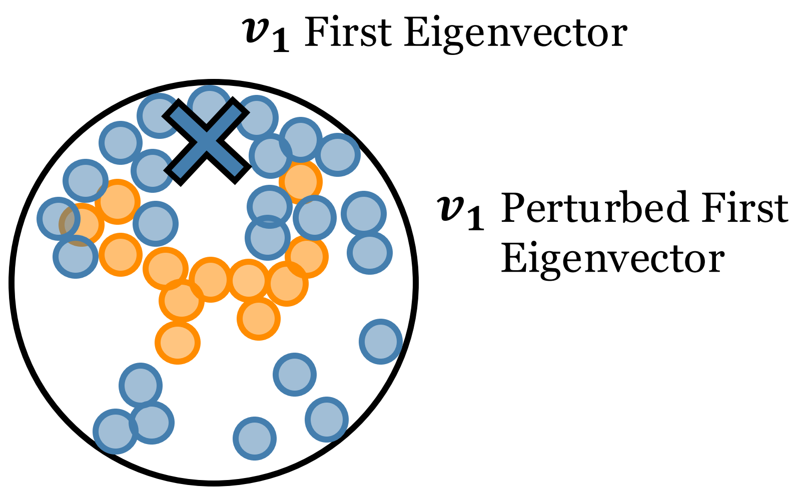


Ideal Spectrum-picking Strategy

Impact of Anomaly Perturbance

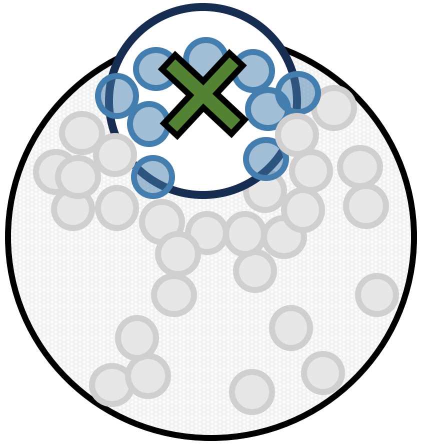


Ideal Spectrum-picking Strategy
Without anomaly perturbation

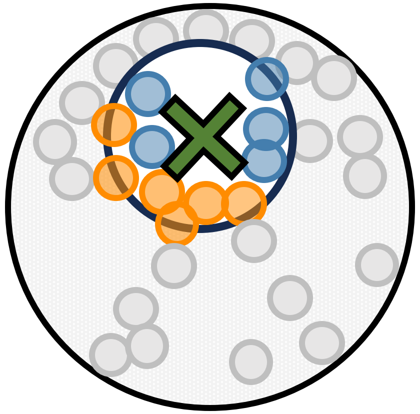


Actual Spectrum-picking Strategy
With anomaly perturbation

Impact of Anomaly Perturbance

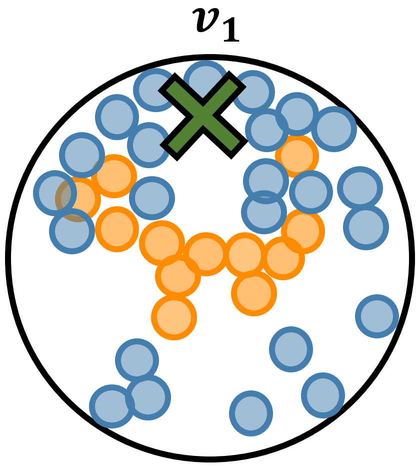


Ideal Spectrum-picking Strategy
Without anomaly perturbation

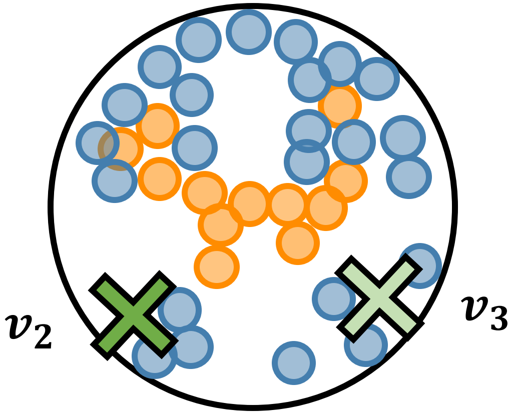


Actual Spectrum-picking Strategy
With anomaly perturbation

Eigenvectors' Vulnerability to Perturbance



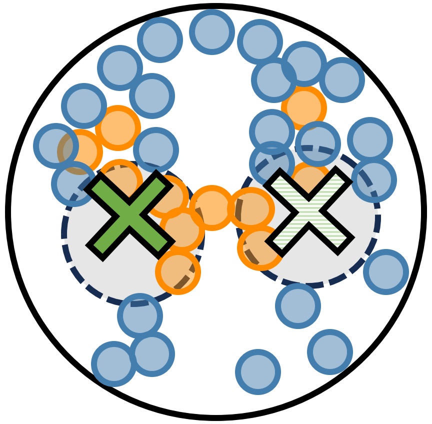
First Eigenvector
Less perturbed



Minor Eigenvectors
More perturbed

- **Intuition**: Minor eigenvectors are more perturbed by anomalies than the major one [91].

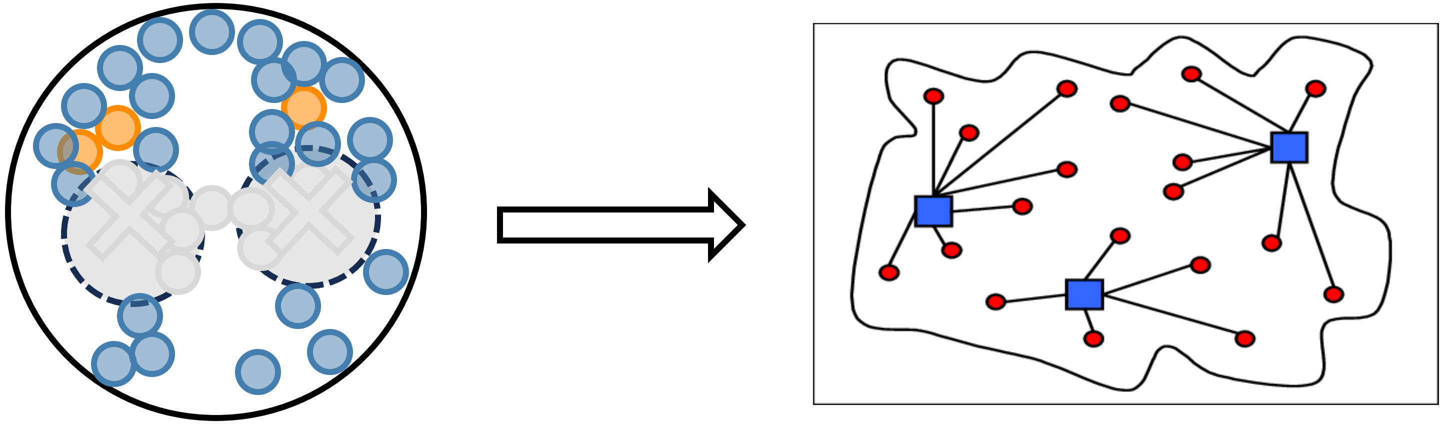
Spectrum-Purifying Sample Selection



Spectrum-purifying Selection Strategy

- **Intuition**: Minor eigenvectors are more perturbed by anomalies than the major one [91].
- **Selection Goal**: discard the samples close to the minor eigenvectors

Spectrum-Purifying Sample Selection



Spectrum-purifying Selection Strategy

Facility Location Problem

- **Intuition**: Minor eigenvectors are more perturbed by anomalies than the major one.
- **Selection Goal**: discard the samples close to the minor eigenvectors
- **Optimization Problem**: K-medoid [1], NP-hard Facility Location Problem solved by greedy with $1-\frac{1}{e}$ approximation.

[1] 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.

Outline

- Background
- Motivation
- Pluto Overview
  - Local Pollution Level Estimation
  - Selection Strategy
  - **Iterative Refinement (Please refer to paper)**
- Experimental Evaluation
- Conclusion

Outline

- Background
- Motivation
- Overview
  - Local Pollution Level Estimation
  - Selection Strategy
  - Iterative Refinement (Please refer to paper)
- **Experimental Evaluation**
- Conclusion

Pollution Impact to Model Training

Dataset	BGL		HDFS		ThunderBird		Spirit	
	Anomaly Ratio ↓	F-1 ↑	Anomaly Ratio ↓	F-1 ↑	Anomaly Ratio ↓	F-1 ↑	Anomaly Ratio ↓	F-1 ↑
Oracle (Clean)	0.0 %	91.69	0.0 %	86.99	0.0 %	90.28	0.0 %	91.69
Original (Polluted)	8.8 %	30.37	2.0 %	24.80	0.8 %	1.70	1.1 %	5.47

Oracle vs. Original: even less than 1% anomalies in the training set can significantly hurt F-1 (90.28 ->1.70).

- F-1 are obtained by training Logbert[1]

[1] Haixuan Guo, Shuhan Yuan, and Xintao Wu. 2021. Logbert: Log anomaly detection via bert. In 2021 International Joint Conference on Neural Networks (IJCNN). IEEE, 1–8.

Pollution Impact to Model Training

Dataset	BGL		HDFS		ThunderBird		Spirit	
	Anomaly Ratio ↓	F-1 ↑	Anomaly Ratio ↓	F-1 ↑	Anomaly Ratio ↓	F-1 ↑	Anomaly Ratio ↓	F-1 ↑
Oracle (Clean)	0.0 %	91.69	0.0 %	86.99	0.0 %	90.28	0.0 %	91.69
Original (Polluted)	8.8 %	30.37	2.0 %	24.80	0.8 %	1.70	1.1 %	5.47
Pluto	0.7 %	62.18	0.0 %	80.59	0.0 %	71.02	0.1 %	51.46

Original vs. Pluto:

- reduces anomaly ratio by >1 orders of magnitude
- to even 0.0 %
- significantly increases F-1 (90.28 ->1.70->71.02).

- F-1 are obtained by training Logbert[1]

[1] Haixuan Guo, Shuhan Yuan, and Xintao Wu. 2021. Logbert: Log anomaly detection via bert. In 2021 International Joint Conference on Neural Networks (IJCNN). IEEE, 1–8.

Overall Performance

Real four log datasets of high-performance computing systems,
all polluted training sets!

	Dataset	BGL				HDFS				ThunderBird				Spirit			
	Metric	P (%)	R (%)	F-1 (%)	Auc	P (%)	R (%)	F-1 (%)	Auc	P (%)	R (%)	F-1 (%)	Auc	P (%)	R (%)	F-1 (%)	Auc
Shallow Log AD	PCA	17.90	11.26	13.82	0.5722	**92.36**	70.31	79.84	**0.9595**	27.12	4.31	7.43	0.8729	8.64	57.28	15.01	0.9398
	IsolationForest	64.99	21.76	32.61	0.8125	44.58	68.62	54.04	0.9341	50.96	6.53	11.57	0.9155	12.44	30.37	17.65	0.9518
	LogCluster	25.93	0.81	1.58	0.5736	96.72	0.62	1.23	0.4104	20.84	0.11	0.22	0.4734	7.75	0.18	0.36	0.7895
	OCSVM	28.42	40.28	33.33	0.7956	8.25	44.30	13.91	0.7908	34.57	15.75	21.64	0.7926	11.46	**85.02**	20.20	0.8498
Deep Log AD	OC4Seq	29.14	61.52	39.55	0.7705	92.19	33.44	49.08	0.7158	**94.15**	52.85	67.7	0.8795	21.14	11.18	14.63	0.7809
	DeepLog	68.63	38.48	49.31	0.8122	57.35	4.13	7.71	0.7029	31.77	4.38	7.71	0.7811	12.75	4.00	6.09	0.9005
	LogBert	73.33	19.15	30.37	0.6861	51.17	16.37	24.8	0.8425	4.62	1.04	1.70	0.9561	8.41	4.16	5.57	0.8995
	LogBert-	**73.95**	19.44	30.79	0.6813	49.16	15.4	23.45	0.8045	4.76	1.04	1.71	0.9585	8.25	4.09	5.47	0.8939
Sample Selection	FINE	44.15	45.33	44.73	0.7674	60.03	35.59	44.69	0.7896	1.20	1.38	1.28	0.8578	4.97	12.03	7.03	0.8853
	ITLM	66.93	24.55	35.92	0.6919	64.73	25.21	36.29	0.8100	5.0	1.38	2.16	0.9523	7.80	4.40	5.63	0.8931
	Co-teaching	61.03	29.72	39.97	0.7048	45.66	46.05	45.85	0.8981	0.44	1.04	0.62	0.9055	6.12	12.45	8.21	0.8996
	PLUTO	55.76	**70.28**	**62.18**	**0.8468**	80.85	**80.34**	**80.59**	0.9496	55.17	**99.65**	**71.02**	**0.9977**	**37.59**	81.56	**51.46**	**0.9623**

Absolute F-1 gain of 17.45% (BGL) to 68.86% (ThunderBird), compared to other sample selection methods.

Thank You & QA

Lei Ma
PhD candidate, WPI
Homepage: https://leima0324.github.io/