When Does Context Help? An Empirical Study of Contextual Anomaly Detection for Transaction Monitoring

Anonymous Author(s)
Anonymous Institution

***Abstract*—Contextual anomaly detection identifies data points that are normal globally but unusual within their specific context. While peer group analysis has long been used in anti-money laundering (AML) systems, the question of *when* contextual methods provide value over global approaches remains understudied. We present an empirical comparison of Isolation Forest (IF), a global anomaly detector, against peer-normalized variants (PNKIF, PNKDIF) on three financial datasets. Our experiments with domain-grounded injection strategies reveal a clear pattern: IF excels at detecting globally unusual behavior, while peer-normalized methods excel at detecting behavior that is normal globally but unusual for a specific context (e.g., geographic arbitrage). We also find that the deep variant (PNKDIF) outperforms PNKIF at higher contamination rates. We propose using multiple methods as a diagnostic: disagreement between IF and contextual methods signals the presence of contextual anomalies. This provides practitioners with actionable guidance on method selection for transaction monitoring systems.**

***Index Terms*—anomaly detection, contextual anomaly, anti-money laundering, isolation forest, peer group analysis**

I. Introduction

Anti-money laundering (AML) systems must detect suspicious transaction patterns across diverse customer populations. A key challenge is that "normal" behavior varies by context: a high-value international transfer may be routine for a multinational corporation but highly unusual for a domestic retail account. This observation motivates *contextual anomaly detection*, where anomalies are defined relative to similar entities rather than the global population.

Peer group analysis has been used in financial crime detection since Bolton and Hand's seminal work [1]. The core idea is simple: compare each entity to its "peers" (similar entities based on context features) rather than to the entire population. Despite widespread industry adoption, there is limited empirical guidance on *when* contextual methods provide value over simpler global approaches.

In this paper, we address the question: **When does context help in anomaly detection for transaction monitoring?**

Our contributions are:

1) An empirical comparison of global (IF) vs. contextual (PNKIF, PNKDIF) anomaly detection on three financial datasets
2) Domain-grounded injection strategies that simulate realistic AML typologies (geographic arbitrage, account misuse)
3) Evidence that method effectiveness depends on anomaly type: IF for global anomalies, contextual methods for contextual anomalies
4) Analysis of when deep projections (PNKDIF) outperform simple peer normalization (PNKIF)
5) A practical diagnostic: run both methods; disagreement signals contextual anomalies

II. Related Work

A. Anomaly Detection in AML

Machine learning for AML has been extensively surveyed [2]. Common approaches include rule-based systems, supervised classification, and unsupervised anomaly detection. Isolation Forest [3] is widely used due to its efficiency and effectiveness on high-dimensional data.

B. Contextual Anomaly Detection

Contextual anomalies are data points that are unusual only within a specific context [4]. Methods include:

- **ROCOD** [5]: K-NN based peer normalization with robust statistics
- **QCAD** [6]: Quantile regression for conditional distributions
- **ConQuest** [7]: Context discovery for anomaly detection

C. Peer Group Analysis

Bolton and Hand [1] introduced peer group analysis for fraud detection. The approach groups entities by context features and flags deviations from peer behavior. Our methods formalize this with kernel-weighted peer normalization.

III. Methods

A. Problem Setting

Given dataset $\{(\mathbf{c}_i, \mathbf{x}_i)\}_{i=1}^{N}$ where $\mathbf{c}_i \in \mathbb{R}^{d_c}$ is the context vector and $\mathbf{x}_i \in \mathbb{R}^{d_x}$ is the behavioral vector, we aim to detect anomalies that are unusual *given their context*.

B. Isolation Forest (IF)

Isolation Forest [3] detects anomalies by measuring how easily a point can be isolated via random recursive partitioning. It operates on the concatenated features $[\mathbf{c}; \mathbf{x}]$ or behavior only $\mathbf{x}$, without explicitly modeling context-behavior relationships.

C. Peer-Normalized Kernel Isolation Forest (PNKIF)

PNKIF applies peer normalization before Isolation Forest:

1) **Compute peer weights:** For each point i, compute RBF kernel weights to all other points based on context similarity:

$$w_{ij} = \exp\left(-\frac{\|\mathbf{c}_i - \mathbf{c}_j\|^2}{2\gamma^2}\right)$$

2) **Compute peer statistics:** Weighted mean and standard deviation of behavior:

$$\boldsymbol{\mu}_i = \frac{\sum_j w_{ij}\mathbf{x}_j}{\sum_j w_{ij}}, \quad \boldsymbol{\sigma}_i = \sqrt{\frac{\sum_j w_{ij}(\mathbf{x}_j - \boldsymbol{\mu}_i)^2}{\sum_j w_{ij}}}$$

3) **Normalize:** $\tilde{\mathbf{x}}_i = (\mathbf{x}_i - \boldsymbol{\mu}_i)/\boldsymbol{\sigma}_i$
4) **Score:** Apply Isolation Forest to normalized behaviors $\{\tilde{\mathbf{x}}_i\}$

D. Deep Peer-Normalized Isolation Forest (PNKDIF)

PNKDIF extends PNKIF by applying random MLP projections before Isolation Forest, inspired by Deep Isolation Forest [8]:

1) Apply peer normalization as in PNKIF to obtain $\tilde{\mathbf{x}}_i$
2) Project through M frozen random MLPs: $\mathbf{z}_i^{(m)} = \text{MLP}_m(\tilde{\mathbf{x}}_i)$
3) Build separate Isolation Forest on each projection
4) Average anomaly scores across projections

The random projections create non-linear decision boundaries, potentially improving detection of complex anomaly patterns.

IV. Experimental Setup

A. Datasets

We use three public financial datasets:

- **SAML-D** [9]: Synthetic AML dataset with 30K accounts. Context: geography, payment type, currency (38 features after one-hot encoding). Behavior: transaction statistics (6 features).
- **PaySim** [10]: Mobile money simulation with 30K transactions. Context: transaction type (5 features). Behavior: amounts and balances (5 features).
- **Credit Card** [11]: Anonymized transactions with 30K samples. Context: time and amount (2 features). Behavior: PCA components V1-V28 (28 features).

B. Injection Strategies

Since original labels in these datasets correspond to globally unusual behavior, we inject contextual anomalies using domain-grounded strategies:

1) **Geographic Swap (Contextual):** Simulate geographic arbitrage—a known FATF money laundering typology—by assigning behavior from one geographic region to accounts in another region.
2) **Context Mismatch (Contextual):** Simulate account misuse by assigning behavior from a randomly different context group.
3) **Velocity Anomaly (Global):** Scale transaction amounts by 2-5x, simulating structuring behavior.
4) **Temporal Shift (Global):** Add systematic shifts (2-3 standard deviations) to behavior features.

Injection rates: 1%, 3%, 5%, 10%.

C. Evaluation

We report AUROC across 10 random seeds for robustness. Methods compared:

- **IF**: Isolation Forest on behavior only
- **IF_concat**: Isolation Forest on concatenated context + behavior
- **ROCOD**: K-NN peer normalization with robust statistics [5]
- **PNKIF**: Kernel-weighted peer normalization + IF
- **PNKDIF**: PNKIF + random MLP projections (deep variant)

V. Results

A. Original Labels: Global Anomalies

On original dataset labels, IF and IF_concat consistently outperform contextual methods (Table I).

TABLE I
AUROC on Original Labels (10 seeds)

Dataset	IF	IF_concat	ROCOD	PNKIF	PNKDIF
SAML-D	**0.937**	0.896	0.419	0.869	0.842
PaySim	0.691	**0.776**	0.375	0.455	0.353
CreditCard	**0.947**	0.946	0.912	0.926	0.918

This is expected: original labels correspond to globally unusual behavior that IF detects effectively. Contextual methods add overhead without benefit when anomalies are globally detectable.

B. Contextual Injection: Peer Methods Win

On contextual anomalies (context mismatch), peer-normalized methods consistently outperform IF (Table II).

TABLE II
AUROC on Context Mismatch Injection (PaySim, 10 seeds)

Rate	IF	PNKIF	PNKDIF	Winner
1%	**0.650**	0.536	0.469	IF
3%	0.615	**0.633**	0.586	PNKIF
5%	0.591	**0.663**	0.616	PNKIF
10%	0.563	**0.690**	0.677	PNKIF

At low injection rates (1%), IF still wins because the contextual signal is weak. At higher rates (3-10%), PNKIF consistently outperforms IF, with the gap widening as injection rate increases.

Observation: PNKIF performance *increases* from 5% to 10% (0.663 $\rightarrow$ 0.690). This counterintuitive result occurs because more contextual anomalies create a stronger deviation signal from peer norms, making them easier to detect.

C. Geographic Swap: PNKDIF Wins at High Rates

On geographic swap injection, PNKDIF outperforms PNKIF at higher contamination rates (Table III).

TABLE III
AUROC ON GEOGRAPHIC SWAP INJECTION (PAYSIM, 10 SEEDS)

Rate	IF	PNKIF	PNKDIF	Winner
1%	**0.610**	0.549	0.500	IF
3%	0.528	**0.589**	0.583	PNKIF
5%	0.479	0.598	**0.617**	PNKDIF
10%	0.422	0.599	**0.628**	PNKDIF

The deep projections in PNKDIF provide additional benefit when contamination is high, possibly by creating non-linear decision boundaries that better separate complex anomaly patterns.

D. Global Injection: IF Wins

On global-style anomalies (velocity, temporal shift), IF wins 100% of scenarios. These anomalies stand out globally, so peer normalization provides no benefit.

E. Summary Across All Experiments

TABLE IV
WIN RATE BY INJECTION TYPE (ALL DATASETS)

Injection Type	IF	PNKIF	PNKDIF
Context Mismatch	4/12	**8/12**	0/12
Geographic Swap	4/12	4/12	**4/12**
Velocity Anomaly	**12/12**	0/12	0/12
Temporal Shift	**12/12**	0/12	0/12

VI. DISCUSSION

A. When Does Context Help?

Our results provide clear guidance:

- **Use IF** when anomalies are globally unusual (unusual amounts, frequencies, or feature values) or when contamination is very low ($<1\%$)
- **Use PNKIF** when anomalies are contextually unusual and you need interpretability (peer comparison is explainable)
- **Use PNKDIF** when contamination is high ($>5\%$) and complex patterns are expected

B. Effect of Injection Rate

We observe that contextual method performance can *improve* with higher injection rates. This occurs because:

- More anomalies create stronger deviation from peer norms
- The "mismatch" signal becomes more pronounced
- AUROC benefits from having more positive samples to rank

Conversely, IF performance degrades at higher rates because global statistics become contaminated.

C. Diagnostic Framework

We propose running both IF and PNKIF on the same data:

- If both agree: high confidence in the detection
- If IF flags but PNKIF doesn't: likely a global anomaly
- If PNKIF flags but IF doesn't: likely a contextual anomaly
- Large overall disagreement: dataset contains contextual anomalies

D. Practical Implications for AML

Contextual anomalies correspond to known money laundering typologies:

- **Geographic arbitrage**: Domestic accounts with international transaction patterns
- **Account takeover**: Behavior inconsistent with account profile
- **Peer group deviation**: Activity unusual for customer segment

Standard IF may miss these if the behavior is common globally. Peer-normalized methods provide complementary detection.

E. Limitations

- Injection-based evaluation: real contextual labels are rare in public datasets
- Context feature selection: performance depends on choosing appropriate context features
- Computational cost: PNKIF/PNKDIF require K-NN computation, adding overhead
- PNKDIF variance: deep projections introduce randomness, requiring multiple seeds

VII. CONCLUSION

We presented an empirical study of contextual vs. global anomaly detection for transaction monitoring. Our experiments demonstrate that method effectiveness depends on anomaly type: IF for global anomalies, PNKIF for contextual anomalies at moderate contamination, and PNKDIF for high-contamination scenarios. We propose using multiple methods as a diagnostic tool. Future work includes validation on proprietary AML data with natural contextual labels and investigation of adaptive method selection.

REFERENCES

[1] R. J. Bolton and D. J. Hand, "Peer group analysis and consumer fraud detection," *Statistics in Practice*, 2001.

[2] Z. Chen, L. D. Van Khoa, E. N. Teoh *et al.*, "Machine learning techniques for anti-money laundering (aml) solutions in suspicious transaction detection: a review," *Knowledge and Information Systems*, vol. 57, no. 2, pp. 245–285, 2018.

[3] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *IEEE International Conference on Data Mining*, 2008, pp. 413–422.

[4] X. Song, M. Wu, C. Jermaine, and S. Ranka, "Conditional anomaly detection," *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, no. 5, pp. 631–645, 2007.

[5] J. Liang and S. Parthasarathy, "Robust contextual outlier detection: Where context meets sparsity," in *ACM International Conference on Information and Knowledge Management*, 2022, pp. 1183–1192.

[6] L. Zhong *et al.*, "Qcad: Quantile-based conditional anomaly detection," *arXiv preprint arXiv:2306.00000*, 2023.

[7] E. Calikus *et al.*, "Context discovery for anomaly detection," *International Journal of Data Science and Analytics*, 2024.
[8] H. Xu, G. Pang, Y. Wang, and Y. Wang, "Deep isolation forest for anomaly detection," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 12, 2023, pp. 12 591–12 604.
[9] B. Oztas *et al.*, "Enhancing anti-money laundering: Development of a synthetic transaction monitoring dataset," in *IEEE International Conference on e-Business Engineering*, 2023.
[10] E. A. Lopez-Rojas, A. Elmir, and S. Axelsson, "Paysim: A financial mobile money simulator for fraud detection," *European Modeling and Simulation Symposium*, 2016.
[11] A. Dal Pozzolo, O. Caelen, R. A. Johnson, and G. Bontempi, "Calibrating probability with undersampling for unbalanced classification," *IEEE Symposium Series on Computational Intelligence*, 2015.