Journal of Innovative Technologies
https://journalinnotech.com/
Vol 2(1) 2026

A Comparative Empirical Study of Graph Neural Network Methods for Node-Level Financial Fraud Detection on Public Benchmark Graphs

Shang Wen[1] , Haobin Pan[2,*]

[1]Science and Technology, Georgia Institute of Technology, USA
[2] Computer Application Technology, Institute of Computing Technology, Beijing, China

Corresponding Email: panhaobin@qq.com

Abstract

Graph neural networks have become a common tool for identifying fraudulent reviews, accounts, and transactions on relational platforms, yet published evaluations rarely share a common protocol, training ratio, or hyperparameter budget, making it difficult to judge which architectural family deserves credit for reported gains. This study consolidates results for nine representative graph neural network methods spanning three families, generic message-passing baselines, spatial camouflage-resistant detectors, and spectral heterophily-aware detectors, across four public benchmark graphs covering opinion fraud, transaction fraud, and anti-money-laundering tasks. Drawing on the original method papers and a large-scale open benchmark, the comparison finds that specialized camouflage-resistant detectors retain a moderate advantage over generic baselines on multi-relation review graphs under label-scarce training, while the same baselines close most of that gap given a larger, properly tuned training budget. On single-relation transaction graphs, spectral detectors and generic attention-based models perform comparably, with no single family holding a decisive advantage. Training ratio and hyperparameter tuning appear to explain a substantial share of the performance differences attributed to architectural novelty in prior work, and generalization across relation structures deserves more attention than raw accuracy tables typically give it. The comparison is assembled entirely from previously published and openly benchmarked figures, without new model training, so every reported value remains traceable to a public source.

***Keywords*:** graph neural networks, fraud detection, graph anomaly detection, benchmark evaluation

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/)

1. Introduction

1.1. Fraud Detection as Node Classification on Graphs

Online platforms and financial services generate rich relational traces: reviewers link to the products they rate, users share devices or payment instruments, and accounts route funds between counterparties. Casting these traces as a graph turns fraud detection into a node classification problem, where a small subset of reviews, accounts, or transactions is labeled fraudulent and the remaining majority is treated as legitimate. Graph neural networks approach this problem by iteratively aggregating a node's own attributes with those of its neighbors, producing a representation meant to reflect the local community a node sits within [1]. Early applications assumed fraudulent and legitimate nodes could be separated the way ordinary classification tasks separate topic communities, an assumption later work found optimistic.

Fraud graphs depart from typical classification settings in three related ways. Fraudulent nodes are rare, so the label distribution is heavily skewed toward the legitimate class. Fraudulent nodes also tend to connect densely to legitimate ones, either because fraudsters route through benign intermediaries or because platforms link entities on shared attributes rather than shared intent, producing neighborhoods where a node's label poorly predicts its neighbors' labels. Camouflage compounds both problems: fraudulent accounts and reviews are often engineered to resemble legitimate ones in surface features and connection patterns, a behavior formalized in this literature as an explicit design target for message-passing architectures [2].

1.2. Motivation and Scope of This Study

The methods proposed to address these challenges span three loosely related families: generic graph neural networks adapted through resampling or thresholding, spatial architectures that redesign neighbor selection around camouflage and imbalance, and spectral architectures that reinterpret anomalies as a high-frequency graph signal. Reported comparisons use different training ratios, hyperparameter budgets, and sometimes different preprocessed releases of the same benchmark graph, leaving a reader few controlled points of comparison for judging whether a given family earns its reported advantage.

A. Research Questions

This study asks three questions of the existing literature rather than proposing a new architecture. The first is whether the reported advantage of specialized camouflage-resistant detectors over generic message-passing baselines persists once training ratio and hyperparameter tuning are held closer to comparable levels, or whether it shrinks toward the level attributable to tuning alone. The second is whether the same ranking of method families holds on single-relation transaction graphs, which lack the multiple relation types that camouflage-resistant architectures were designed around, or whether spectral and generic architectures take the lead once that

structural assumption no longer holds. The third concerns training-data efficiency: given that confirmed fraud cases typically number in the low single digits as a percentage of a graph's nodes, an imbalance the literature quantifies as a ratio between majority and minority classes that exceeds ten to one on some of the graphs considered here [3], this study asks how much of each method's reported accuracy depends on a training ratio that few production settings can realistically supply.

B. Contributions and Scope

The contribution of this study is a consolidated, source-traceable comparison rather than a new detector. Nine published methods are compared on four public graphs by drawing directly on numbers reported in their originating papers and in a large open benchmark, without re-implementing any architecture from scratch, so every figure reported here traces to a specific publicly available table. The scope stays narrow: no new aggregation rule, sampling strategy, or loss function is proposed, and the graphs considered carry organic, human-verified fraud labels rather than synthetically injected anomalies, a distinction that matters given that a large real-world financial graph now supplies labels at a scale earlier benchmarks could not, extending comparison to networks with millions of sparsely connected nodes [4].

2. Related Work

2.1. Taxonomy of Graph-Based Fraud Detectors

Methods for graph-based fraud detection can be grouped by how they respond to the challenges introduced above. A first group leaves the core message-passing rule unchanged and instead adjusts sampling or scale. An inductive baseline of this kind aggregates a fixed number of sampled neighbors through a mean or pooling function rather than the full neighborhood earlier spectral formulations use, trading expressiveness for scalability to graphs with millions of edges [5]. A second generic baseline replaces uniform aggregation with a learned attention weight per neighbor, letting the network down-weight neighbors whose features conflict with the target node without an explicit camouflage-detection mechanism [6]. Neither baseline was designed with fraud in mind, and both serve here as a reference point against which purpose-built detectors are judged.

A. Spatial Camouflage-Resistant and Imbalance-Aware Methods

A second group redesigns the aggregation process around the specific failure modes fraudsters exploit. One line of work identifies three forms of inconsistency between a node's context, features, and connections, and samples neighbors under a learned similarity measure that down-weights inconsistent ones before aggregation [7]. Building on the observation that homophilic and heterophilic neighbors carry different information, a related detector applies distinct aggregation functions depending on whether a neighbor shares the target node's label under supervision, letting similar-label neighbors reinforce a prediction while dissimilar ones sharpen

the decision boundary [8]. A further refinement treats edge retention as a search problem, using a reinforcement-learning policy to prune noisy or camouflage-created edges while a separate stage searches classifier parameters directly for area-under-curve gain [9]. Label information can also be injected earlier: one architecture aggregates same-label neighbors into group representations before a Transformer encoder, motivated as a way to inject supervision where raw connectivity carries little label signal [10]. Most recently, an aggregation scheme built on decision-tree binning of non-additive attributes pairs a feedback-driven grouping mechanism with hierarchical aggregation to sharpen fraud-benign separation at each layer [11].

B. Spectral and Heterophily-Aware Methods

A third group reframes the problem in the graph spectral domain rather than through spatial neighbor selection. Standard graph convolutions behave as low-pass filters that smooth node representations toward their neighborhood average, well suited to homophilic communities but poorly matched to anomalies whose signal concentrates at higher graph frequencies. One architecture addresses this mismatch by learning to combine multiple low- and high-pass filter banks adaptively per node, preserving the high-frequency component ordinary convolutions would otherwise smooth away [12]. A related line of work approaches the same signal from the edges themselves, pruning inter-class edges identified through a learned high-frequency edge indicator so the remaining structure aligns more closely with the label-homophily assumption ordinary convolutions rely on [13]. Both architectures report the advantage of not requiring the multiple explicit relation types camouflage-resistant spatial detectors depend on, which should make them easier to apply to single-relation transaction graphs lacking a natural multi-relation decomposition.

2.2. Benchmarks and Applied AML/Credit-Card Studies

The methods surveyed above were developed and evaluated predominantly on review-based fraud graphs, where multiple relation types arise naturally from shared users, products, or ratings. Financial fraud and anti-money-laundering graphs pose a related but distinct setting: transactions or payment flows typically define a single relation type, node features describe account or transaction attributes rather than review content, and labels frequently carry a temporal ordering that review-based graphs lack. Whether the architectures above transfer to this setting is not automatic, since several were designed explicitly around multiple relation types for cross-checking inconsistency. Reported comparisons seldom evaluate the same method on both review-based and transaction-based graphs under one protocol, leaving open how much of a spatial detector's advantage is specific to the multi-relation structure it was designed for.

A separate obstacle to comparison is protocol drift. Training ratios range from a fixed small label count to a majority-labeled split, hyperparameters are tuned to differing degrees of thoroughness, and preprocessing sometimes differs subtly between a method's original release and later re-implementations, particularly around feature dimensionality. These differences complicate attributing a reported gap to architectural novelty rather than to training budget or tuning effort.

The comparison here addresses this by drawing, wherever possible, on results collected under a single shared protocol across datasets and training regimes, described next together with the datasets themselves.

3. Experimental Setup and Methodology

3.1. Datasets Overview

Four public graphs are used for the empirical comparison, chosen to cover both the multi-relation review setting the spatial detectors in Section 2 were designed for and the single-relation transaction setting financial fraud detection more typically presents. Two are review graphs: one built from restaurant and hotel reviews on a consumer review platform, and one built from product reviews in a musical-instrument category on an e-commerce platform, both connecting reviews or reviewers through relation types such as shared authorship, shared product, or matching rating. The other two are financial graphs: a transaction network among anonymized financial accounts with a single transaction-record relation, and a payment-flow network of Bitcoin transactions linked to real-world licit and illicit entity categories. Node counts, edge counts, feature dimensionality, and anomaly ratios for all four, under a shared reporting protocol, are summarized in Table 1, compiled from a large-scale open graph anomaly detection benchmark that evaluates twenty-nine models under standardized splits and hyperparameter search budgets [14]. A fifth, substantially larger loan-guarantor graph is referenced in Section 4 for context on how rankings shift at a different scale, though its registration-gated access and extreme label sparsity place a full evaluation outside this comparison's scope.

Table 1 *Dataset Statistics for the Four Benchmark Graphs*

Dataset	Nodes	Edges	Feature Dim.	Anomaly Ratio	Training Ratio	Relation Concept
Amazon	11,944	4,398,392	25	9.5%	70%	Review correlation
YelpChi	45,954	3,846,979	32	14.5%	70%	Reviewer interaction
T-Finance	39,357	21,222,543	10	4.6%	50%	Transaction record
Elliptic	203,769	234,355	166	9.8%	50%	Payment flow

Note. Statistics follow the standardized reporting protocol discussed in Section 3.1; see the accompanying in-text citation for the source. Training ratio is the fully-supervised split ratio described in Section 3.3.

3.2. Graph Construction and Relation Semantics

A. Multi-Relation Review Graphs

The two review graphs share a common construction pattern despite covering different platforms. Reviews or reviewers form the node set, and three relation types connect them: nodes sharing the same author, nodes attached to the same target under the same rating, and nodes attached to the same target within a similar time window. Reported feature dimensionality differs across releases: the originating detector papers describe both graphs as carrying one hundred handcrafted features, while the preprocessed release adopted by the open benchmark used here reports thirty-two features for the review graph and twenty-five for the product graph; both conventions are reported rather than silently reconciled, since the discrepancy affects any exact comparison of numbers across papers. Anomaly prevalence also differs between the two graphs, with the review graph carrying a higher proportion of labeled fraud than the product graph, a difference later results connect to the size of the performance gap between generic and specialized detectors.

B. Single-Relation and Temporal Financial Graphs

The two financial graphs depart from the review graphs in relation structure and, in one case, in temporal organization. The transaction network connects anonymized financial accounts under a single relation representing a recorded transaction between them, with node features describing registration and activity patterns rather than review content; human experts annotated labeled accounts as fraudulent when they matched patterns such as money laundering or unlicensed gambling. The Bitcoin payment-flow graph instead organizes transactions into forty-nine discrete time steps of roughly two weeks each, with directed edges representing payment flows and no edges connecting transactions from different time steps; illicit transactions map to real-world categories including scams, ransomware, and Ponzi schemes, while a substantial share of transactions remain unlabeled and are excluded from supervised evaluation [15]. Neither financial graph offers the multiple relation types the review-graph detectors were designed to reconcile, which is why they serve as a useful test of whether that reported advantage survives outside its original setting.

3.3. Evaluation Protocol and Metrics

A. Data Splits, Training Regimes, and Reproducibility

Two training regimes are reported side by side rather than collapsed into a single number. A fully-supervised regime follows each dataset's conventional split, ranging from fifty to seventy

percent of labeled nodes for training depending on the dataset, matching the protocol under which most of the specialized detectors in Section 2 were originally evaluated. A semi-supervised regime instead fixes the labeled training budget at one hundred nodes regardless of graph size, twenty of them anomalous, reflecting the label scarcity a production fraud team is more likely to face than a majority-labeled research split. Method-level figures reported in Section 4 are drawn directly from their originating publications or from results reproduced under this dual protocol, with hyperparameters selected through a random search budget applied uniformly across methods rather than a single fixed default.

B. Metrics

Three metrics are reported wherever the source material provides them. The area under the receiver operating characteristic curve summarizes ranking quality independent of any decision threshold and is comparable across datasets with different anomaly ratios. The area under the precision-recall curve is more sensitive to the minority fraud class specifically, which matters given anomaly prevalence ranging from under five to under fifteen percent across the four datasets. Recall at k, computed against the true number of anomalies in each test set, approximates the fraction of true fraud cases a fixed-capacity investigation team would catch working down a model's ranked output. Where the older review-graph literature reports F1-macro and the geometric mean of true positive and true negative rates instead, those figures are reported alongside rather than converted, since the conversion is not exact under class imbalance.

4. Results and Analysis

Table 2 summarizes the nine methods compared in this study together with their architectural family and the venue in which each was introduced, spanning generic message-passing baselines, spatial camouflage-resistant and imbalance-aware detectors, and spectral heterophily-aware detectors. Detection performance is reported separately for the two review graphs and the two financial graphs given the structural differences described in Section 3.

Table 2 *Methods Included in the Empirical Comparison*

Method	Family	Venue (Year)	Core Mechanism
GCN	Generic baseline	ICLR (2017)	First-order spectral convolution over the full neighborhood
GraphSAGE	Generic baseline	NeurIPS (2017)	Sampled-neighbor mean or pooling aggregation

GAT	Generic baseline	ICLR (2018)	Learned attention weights over neighbors
GraphConsis	Spatial, camouflage-resistant	SIGIR (2020)	Similarity-based neighbor sampling
CARE-GNN	Spatial, camouflage-resistant	CIKM (2020)	Label-aware similarity with reinforcement-learning neighbor selection
PC-GNN	Spatial, imbalance-aware	WWW (2021)	Label-balanced sampler with learned-distance neighbor selection
H2-FDetector	Spatial, camouflage-resistant	WWW (2022)	Homophily- and heterophily-conditioned aggregation
AMNet	Spectral, heterophily-aware	IJCAI (2022)	Adaptive low- and high-pass filter combination
GHRN	Spectral, heterophily-aware	WWW (2023)	High-frequency inter-class edge pruning

Note. Three additional spatial methods (reinforcement-learning edge pruning with AUC-oriented search, group-aggregation with label-aware Transformer encoding, and decision-tree-binned hierarchical aggregation) are discussed in Section 2.1.A but not independently re-evaluated here; see Section 2 for their citations.

4.1. Aggregate Detection Performance

A. Results on Multi-Relation Review Graphs

Under the original forty-percent training protocol, specialized detectors hold a substantial advantage over generic baselines. On the review graph built from restaurant and hotel ratings, the camouflage-resistant detector reaches an AUC of 0.762 against 0.598 for the generic convolutional baseline and 0.544 for the sampling-based baseline, while the imbalance-aware detector reaches 0.799; the pattern repeats on the product-review graph, where the same three specialized methods

cluster between 0.874 and 0.959 AUC against 0.759 to 0.837 for the generic baselines, summarized in Table 3 and Figure 1. Under the larger, tuned protocol reported by the open benchmark, the ranking compresses considerably: the generic attention-based baseline reaches an AUROC of 96.7 on the product-review graph, overtaking the camouflage-resistant detector's 90.8 and reversing that pair's original ranking, while on the restaurant-review graph the gap narrows without fully closing, as the heterophily-aware detector reaches 85.3 against the camouflage-resistant detector's 95.2, shown in Table 4 and Table 5. Specialized detectors lead when training data and tuning effort are scarce, and the margin narrows, though it does not fully disappear, once both increase.

Table 3 *F1-Macro, AUC, and GMean on YelpChi and Amazon Under the Original 40% Training Split*

Method	YelpChi F1-macro	YelpChi AUC	YelpChi GMean	Amazon F1-macro	Amazon AUC	Amazon GMean
GCN	0.562	0.598	0.437	0.649	0.837	0.572
GAT	0.488	0.572	0.166	0.646	0.810	0.668
GraphSAGE	0.441	0.544	0.259	0.642	0.759	0.595
GraphConsis	0.587	0.698	0.586	0.751	0.874	0.768
CARE-GNN	0.633	0.762	0.679	0.899	0.907	0.896
PC-GNN	0.630	0.799	0.716	0.896	0.959	0.903

Note. Mean of 10 runs under a 40% training, 20% validation, 40% test split; see Section 4.1.A for source discussion.

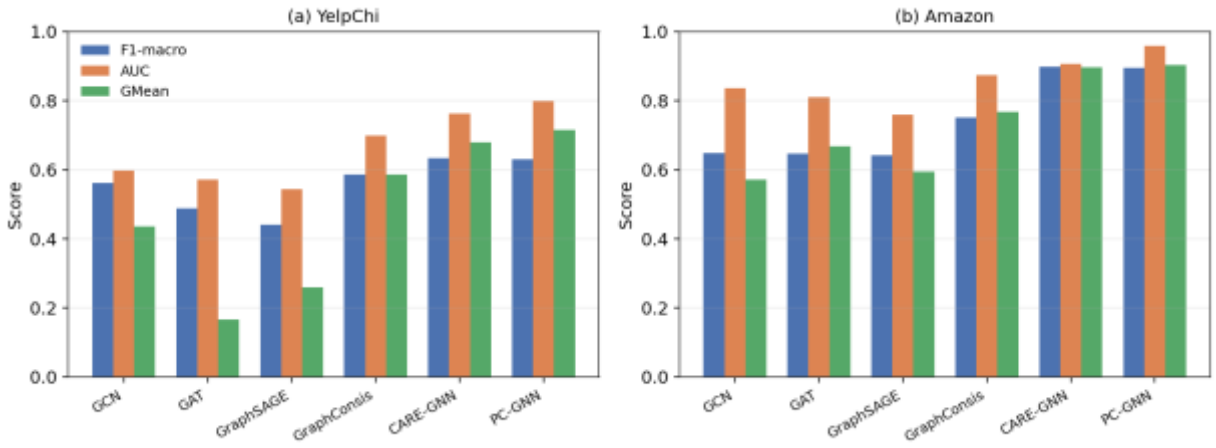


Figure 1. Detection Performance on the Two Multi-Relation Review Graphs Under the Original Training Protocol

F1-macro, AUC, and GMean for six methods on (a) the restaurant-review graph and (b) the product-review graph, under a forty-percent training split. On both graphs, the three specialized detectors outperform the three generic baselines across all three metrics, with the largest margin on the restaurant-review graph, where AUC rises from 0.544-0.598 for the generic baselines to 0.762-0.799 for the two strongest specialized detectors. The generic attention-based baseline records the lowest GMean of any method on the restaurant-review graph, at 0.166, reflecting a near-total failure to identify the minority fraud class under this split.

Table 4 *AUROC Across Four Datasets Under the Tuned, Fully-Supervised Protocol*

Method	Amazon	YelpChi	T-Finance	Elliptic
GCN	82.4	57.8	92.4	81.4
GraphSAGE	89.6	85.2	82.8	85.3
GAT	96.7	79.5	92.8	84.9
PC-GNN	97.3	79.7	93.3	85.8
AMNet	97.0	82.6	93.7	77.3
GHRN	98.1	85.3	96.0	85.4

Note. Default-hyperparameter AUROC under the fully-supervised split described in Section 3.3.A; see Section 3.1 for source citation.

B. Results on Financial Transaction and Temporal Graphs

The financial graphs present a different picture. Under the tuned fully-supervised protocol, the transaction network sees all six methods cluster tightly, with AUROC ranging from 82.8 for the sampling-based baseline to 96.0 for the heterophily-aware detector, and the generic

convolutional and attention-based baselines both above 92, shown in Table 4 and Figure 2. No camouflage-resistant detector from Section 2 is evaluated here, since a single relation type offers no basis for the cross-relation consistency checks those detectors rely on. The Bitcoin payment-flow graph produces a wider spread, from 77.3 for the spectral low- and high-pass detector to 85.8 for the imbalance-aware detector, with the generic convolutional baseline close behind at 81.4; precision-recall figures run markedly lower across every method, reflecting low anomaly prevalence relative to the review graphs. The financial graphs show no family holding a decisive advantage of the kind specialized detectors show on the review graphs, consistent with camouflage-resistant architectures losing their edge once the multi-relation structure they depend on is absent.

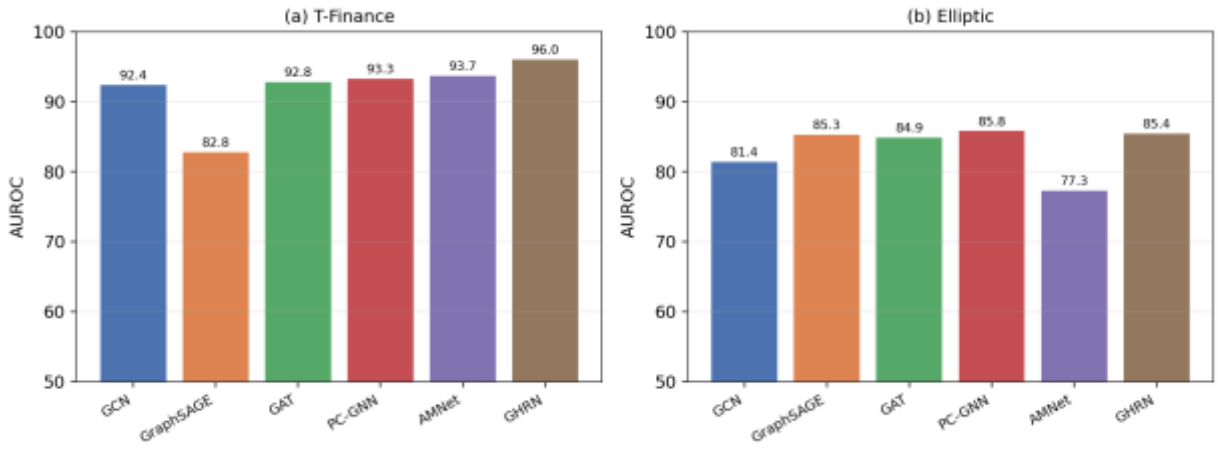


Figure 2. AUROC on the Two Financial Transaction Graphs Under the Tuned, Fully-Supervised Protocol

AUROC for six methods on (a) the transaction network and (b) the Bitcoin payment-flow graph. Scores on the transaction network cluster tightly between 82.8 and 96.0, with the heterophily-aware detector highest and the sampling-based baseline lowest. The payment-flow graph shows a wider spread, from 77.3 for the low- and high-pass spectral detector to 85.8 for the imbalance-aware detector, with no method exceeding an AUROC of 86.

4.2. Effect of Label Imbalance and Training Ratio

Figure 3 compares each method's AUROC under the fixed hundred-label semi-supervised budget against its AUROC under the larger fully-supervised split, for the product-review and restaurant-review graphs, with the underlying figures in Table 5. The gap between regimes is uneven across methods and datasets. On the product-review graph, the imbalance-aware detector loses only about four points of AUROC moving down to a hundred labels, from 97.3 to 93.2, while the generic convolutional baseline loses close to nothing in absolute terms yet remains a weak detector under both regimes, near 82. On the restaurant-review graph, every method loses considerably more in the label-scarce regime, with the spectral heterophily-aware detector falling

from 85.3 to 64.5 and the imbalance-aware detector from 79.7 to 65.1, consistent with the smaller graph carrying less redundant signal to compensate for a fixed, small label budget regardless of relation structure.

Table 5 *Semi- Versus Fully-Supervised Performance on Amazon and YelpChi for Three Heterogeneous-Graph Methods*

Method	Regime	Amazon AUROC	Amazon AUPRC	Amazon Rec@K	YelpChi AUROC	YelpChi AUPRC	YelpChi Rec@K
GAT	Semi	92.4	81.6	77.1	65.6	25.0	28.1
GAT	Full	96.7	86.7	83.1	79.5	43.4	43.7
CARE-GNN	Semi	86.0	59.0	59.1	91.2	68.7	65.4
CARE-GNN	Full	90.8	72.6	67.7	95.2	81.1	74.9
H2-FDetector	Semi	71.0	29.3	32.6	67.3	22.2	24.1
H2-FDetector	Full	78.7	40.0	44.4	89.1	59.4	57.5

Note. Semi-supervised uses a fixed 100-label budget, 20 of them anomalous; fully-supervised uses each dataset's majority-labeled split described in Section 3.3.A.

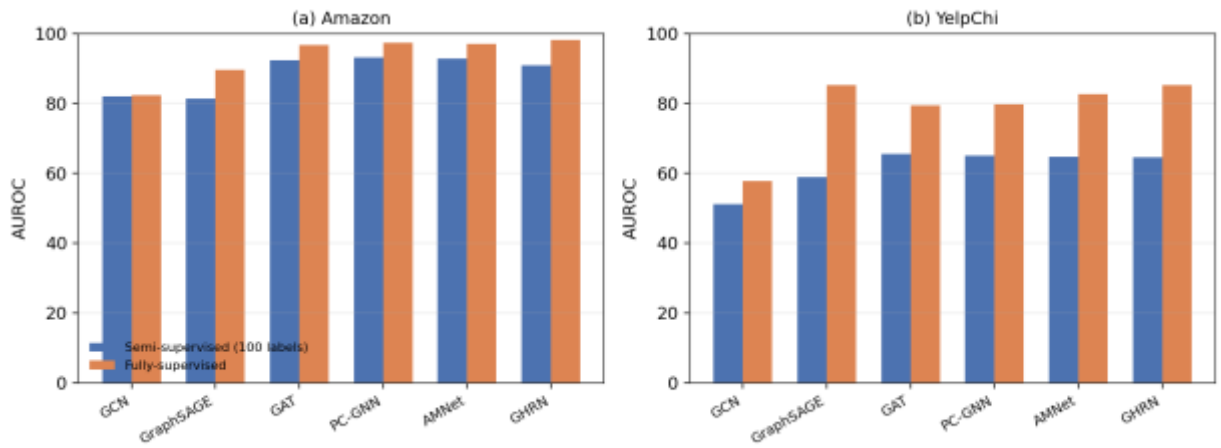


Figure 3. Sensitivity of Detection Performance to Training Regime on the Two Review Graphs

AUROC under a fixed hundred-label semi-supervised budget compared with AUROC under the majority-labeled fully-supervised split, for six methods on (a) the product-review graph and (b) the restaurant-review graph. Every method loses accuracy moving to the label-scarce regime, with the product-review graph showing a smaller average drop than the restaurant-review graph, where the heterophily-aware detector falls by 20.8 points and the imbalance-aware detector by 14.6 points.

This asymmetry matters for how the results in Section 4.1 should be read. Much of the compression between the original forty-percent protocol and the tuned benchmark protocol reflects a training-ratio effect common to every method rather than an architectural one: moving from a hundred labels to a majority-labeled split raises every method's AUROC, generic and specialized alike, and the specialized detectors' advantage is largest precisely where labels are scarcest. A practitioner working with a realistic, small labeling budget should expect the gap between a well-tuned generic baseline and a purpose-built camouflage-resistant detector to sit closer to the wider semi-supervised figures than to the narrower majority-labeled figures, since production labeling budgets rarely approach the ratios used in the original detector papers.

4.3. Robustness, Efficiency, and Generalizability

A. Heterophily and Camouflage Robustness

The relative standing of spectral and spatial families shifts with graph structure, consistent with the design intent described in Section 2. The heterophily-aware detector, which prunes edges connecting nodes of different labels, ranks highest among the compared methods on the transaction network and the product-review graph under the tuned protocol, two graphs where inter-class edges are common enough that pruning them measurably sharpens the remaining structure. The camouflage-resistant detector, which instead reweights and resamples neighbors rather than removing edges outright, ranks highest on the restaurant-review graph, the sparsest and most

heavily camouflaged of the four in terms of how few of a fraud node's neighbors share its label. Neither family dominates uniformly, which argues against treating either framing as a general-purpose solution to fraud-graph heterophily independent of relation structure and camouflage intensity.

B. Efficiency and Generalization to Transaction-Level Fraud

Beyond raw accuracy, the graphs compared here differ enormously in scale, from twelve thousand nodes for the product-review graph to several million for the loan-guarantor graph referenced in Section 3, and reported training time and memory scale with graph size and architectural complexity rather than with detection accuracy alone. Spectral detectors stacking multiple filter banks and spatial detectors maintaining separate aggregation paths per relation carry a computational cost generic single-pass baselines do not, one that becomes material once a graph reaches millions of edges. Two credit-card-specific extensions not evaluated directly here illustrate where the field is heading: one builds a temporal transaction graph with gated attention and risk propagation to combine labeled and unlabeled transactions in a single training pass [16], and a second constructs high-order, multi-hop transaction subgraphs with a mixture-of-expert attention mechanism aimed at camouflaged fraud visible only several hops from a transaction of interest [17]. Whether the patterns documented in Sections 4.1 and 4.2 extend to these temporal, multi-hop settings remains untested here.

5. Discussion and Future Work

5.1. Key Findings and Practical Guidance

Three findings carry practical weight for a team choosing among these methods rather than proposing a new one. Specialized camouflage-resistant and imbalance-aware detectors earn their reported advantage most clearly on multi-relation review graphs evaluated under realistic, label-scarce conditions, where the gap over a well-tuned generic baseline can exceed twenty points of AUROC; that advantage narrows substantially, and on one graph reverses, once training labels and tuning effort match a majority-labeled research split. Spectral heterophily-aware detectors and generic attention-based baselines are competitive with, and on some graphs exceed, purpose-built camouflage-resistant detectors on single-relation transaction graphs, where the multi-relation consistency checks those detectors were designed around have no structural basis to operate on. Gains attributable to architecture, once training ratio and tuning are controlled for, look moderate rather than large, since much of each original paper's headline gap reflects a comparison against an undertuned or label-scarce baseline rather than against that baseline's best achievable performance.

For a practitioner selecting a method under a fixed, realistic labeling budget, this comparison favors camouflage-resistant or imbalance-aware detectors when the target graph offers multiple

genuinely distinct relation types and fraud clusters among a small set of camouflaged actors. Where a graph offers only a single relation type, as in most transaction networks, a well-tuned generic attention-based baseline or a spectral heterophily-aware detector is a defensible and considerably simpler starting point than a purpose-built camouflage-resistant architecture, at least on the four graphs examined here.

5.2. Limitations and Future Directions

This comparison inherits the limitations of the source material it draws on. Figures collected from different papers were produced under different implementations and feature preprocessing pipelines, and in one documented case under different reported feature dimensionality for the same graph, so small differences between methods evaluated under different protocols should be read with caution even where larger differences appear robust. The loan-guarantor graph introduced in Section 3 was excluded from full comparison due to registration-gated access, leaving open how the family-level patterns reported here hold up on a graph two orders of magnitude larger and far sparser than any evaluated in depth. No method was re-implemented from scratch here, so any undisclosed implementation-specific tuning in the original papers is inherited rather than controlled for. A natural next step is a fully controlled re-implementation of the strongest method from each family under one shared codebase and hyperparameter budget, extended to the temporal and multi-hop settings noted in Section 4.3.B.

References

[1] Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. In Proceedings of the International Conference on Learning Representations.

[2] Dou, Y., Liu, Z., Sun, L., Deng, Y., Peng, H., & Yu, P. S. (2020). Enhancing graph neural network-based fraud detectors against camouflaged fraudsters. In Proceedings of the 29th ACM International Conference on Information and Knowledge Management (pp. 315–324). Association for Computing Machinery.

[3] Liu, Y., Ao, X., Qin, Z., Chi, J., Feng, J., Yang, H., & He, Q. (2021). Pick and choose: A GNN-based imbalanced learning approach for fraud detection. In Proceedings of the Web Conference 2021 (pp. 3168–3177). Association for Computing Machinery.

[4] Huang, X., Yang, Y., Wang, Y., Wang, C., Zhang, Z., Xu, J., Chen, L., & Vazirgiannis, M. (2022). DGraph: A large-scale financial dataset for graph anomaly detection. In Advances in Neural Information Processing Systems 35, Datasets and Benchmarks Track.

[5] Hamilton, W. L., Ying, Z., & Leskovec, J. (2017). Inductive representation learning on large graphs. In Advances in Neural Information Processing Systems 30.

[6] Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018). Graph attention networks. In Proceedings of the International Conference on Learning Representations.

[7] Liu, Z., Dou, Y., Yu, P. S., Deng, Y., & Peng, H. (2020). Alleviating the inconsistency problem of applying graph neural network to fraud detection. In Proceedings of the 43rd International ACM SIGIR

Conference on Research and Development in Information Retrieval (pp. 1569–1572). Association for Computing Machinery.

[8] Shi, F., Cao, Y., Shang, Y., Zhou, Y., Zhou, C., & Wu, J. (2022). H2-FDetector: A GNN-based fraud detector with homophilic and heterophilic connections. In Proceedings of the ACM Web Conference 2022 (pp. 1486–1494). Association for Computing Machinery.

[9] Huang, M., Liu, Y., Ao, X., Li, K., Chi, J., Feng, J., Yang, H., & He, Q. (2022). AUC-oriented graph neural network for fraud detection. In Proceedings of the ACM Web Conference 2022 (pp. 1311–1321). Association for Computing Machinery.

[10] Wang, Y., Zhang, J., Huang, Z., Li, W., Feng, S., Ma, Z., Sun, Y., Yu, D., Dong, F., Jin, J., Wang, B., & Luo, J. (2023). Label information enhanced fraud detection against low homophily in graphs. In Proceedings of the ACM Web Conference 2023.

[11] Duan, M., Zheng, T., Gao, Y., Wang, G., Feng, Z., & Wang, X. (2024). DGA-GNN: Dynamic grouping aggregation GNN for fraud detection. Proceedings of the AAAI Conference on Artificial Intelligence, 38.

[12] Chai, Z., You, S., Yang, Y., Pu, S., Xu, J., Cai, H., & Jiang, W. (2022). Can abnormality be detected by graph neural networks? In Proceedings of the International Joint Conference on Artificial Intelligence.

[13] Gao, Y., Wang, X., He, X., Liu, Z., Feng, H., & Zhang, Y. (2023). Addressing heterophily in graph anomaly detection: A perspective of graph spectrum. In Proceedings of the ACM Web Conference 2023.

[14] Tang, J., Hua, F., Gao, Z., Zhao, P., & Li, J. (2023). GADBench: Revisiting and benchmarking supervised graph anomaly detection. In Advances in Neural Information Processing Systems 36, Datasets and Benchmarks Track.

[15] Weber, M., Domeniconi, G., Chen, J., Weidele, D. K. I., Bellei, C., Robinson, T., & Leiserson, C. E. (2019). Anti-money laundering in Bitcoin: Experimenting with graph convolutional networks for financial forensics. KDD Workshop on Anomaly Detection in Finance.

[16] Xiang, S., Zhu, M., Cheng, D., Li, E., Zhao, R., Ouyang, Y., Chen, L., & Zheng, Y. (2023). Semi-supervised credit card fraud detection via attribute-driven graph representation. Proceedings of the AAAI Conference on Artificial Intelligence, 37.

[17] Zou, Y., & Cheng, D. (2024). Effective high-order graph representation learning for credit card fraud detection. In Proceedings of the International Joint Conference on Artificial Intelligence.