EnGraph: Ensemble-Based Augmentation for Graph Anomaly Detection

Andrew Shields
Munster Technological University
Kerry, Ireland
andrew.shields@mtu.ie

Pat Doody
Munster Technological University
Kerry, Ireland
pat.doody@mtu.ie

Robert Sheehy
Munster Technological University
Kerry, Ireland
robert.sheehy@mtu.ie

ABSTRACT

Graph anomaly detection is critical for identifying rare, irregular patterns in complex networks, particularly when labelled data is scarce and class imbalance is high. This paper introduces EnGraph, a novel ensemble-based framework that combines ComplEx knowledge graph embeddings, pseudo-labelling, and targeted data augmentation to improve anomaly detection in attributed graphs.

EnGraph integrates multiple unsupervised base models and leverages a generative adversarial network (GAN) to synthesise high-confidence anomalous embeddings, which are iteratively used to refine pseudo-labels. Evaluated across 14 datasets, synthetic, real-world, and injected anomaly benchmarks,

EnGraph performs comparably to supervised methods while requiring fewer labelled instances. Performance in low-label and imbalanced scenarios. Notably, it achieves up to 12% higher AUC-PR and 10% higher F1-score compared to leading baselines on datasets with <5% anomaly ratios. Results also show EnGraph's robustness across varying graph sizes and structures, with competitive AUC-ROC scores maintained even in extreme sparsity. The method is reproducible, scalable, and applicable to diverse graph domains such as e-commerce, citation networks, and social platforms.

PVLDB Reference Format:

Andrew Shields, Pat Doody and Robert Sheehy EnGraph: Ensemble-Based Augmentation for Graph Anomaly Detection. VLDB 2025 Workshop: 4th International Workshop on Large-Scale Graph Data Analytics (LSGDA 2025).
doi:XX.XX/XXX.XX

PVLDB Artifact Availability:

The source code, data, and/or other artifacts have been made available at http://vldb.org/pvldb/format_vol14.html.

Proceedings of the VLDB Endowment, Vol. 14, No. 1 ISSN 2150-8097.
doi:XX.XX/XXX.XX

1 INTRODUCTION

Graph-based anomaly detection presents unique challenges due to the interlinked nature of structural and attribute information, the rarity of anomalies, and the lack of annotated data [1], [2]. These characteristics reduce the effectiveness of traditional anomaly detection methods that assume feature independence or rely on large, labelled datasets. Recent work has shifted towards unsupervised and semi-supervised models [3], [4], [5], [6], but scalability, class imbalance, and robustness to graph perturbation remain open issues [7], [8].

Recent developments in unsupervised graph learning, particularly through graph neural networks (GNNs) and autoencoders, have advanced the field by enabling scalable anomaly detection without requiring dense supervision [9], [4], [8]. Notable contributions include GCN-based anomaly detectors, such as DOMINANT [8], and autoencoder frameworks like GCNAE [3], as well as contrastive representation learning methods, including AdONE and CoLA [9], [7]. While these methods perform well under certain conditions, they are frequently sensitive to the extreme class imbalance inherent in real-world datasets and often require retraining for each new domain or graph instance [1].

Many unsupervised and semi-supervised models assume that anomalies have significant deviations from normal data distributions [2]. [10]. However, graph anomalies can be subtle, sparse, or context-dependent, complicating detection [1], [8]. Recent advances, such as graph contrastive learning GADFormer [11] and transformer-based models GTN [12], as well as ensemble-based outlier detectors XGBOD [13], have sought to overcome these limitations by integrating representation learning with scalable detection frameworks, often in a semi-supervised or hybrid setting. These approaches demonstrate improved robustness to imbalance, better feature diversity, and enhanced anomaly recall, particularly when labelled data is scarce [3].

Despite these advances, an important gap remains in methods that can effectively generalise to unseen graph data while maintaining detection robustness under imbalanced conditions [6], [7], [8]. Data augmentation with ensemble learning may improve model generalisability, promote diverse decision boundaries, and mitigate overfitting. However, it remains under-explored in GAD research.

This paper introduces EnGraph, a framework that combines knowledge graph embeddings (ComplEx [14]), pseudo-labelling, and ensemble learning to address the challenges of sparse labels and class imbalance in graph anomaly detection. Unlike prior work, EnGraph directly integrates GAN-based synthetic augmentation into an iterative pseudo-labelling loop and aggregates anomaly scores from structurally diverse detectors. This builds on prior studies using knowledge graph embeddings, such as TransE [15], RotatE [16], and BESS [17], which have demonstrated the potential of embedding-based models for relational anomaly detection. We evaluate this framework across 14 graph datasets, including temporal and injected anomaly benchmarks, demonstrating improved robustness and scalability under diverse anomaly conditions.

2 BACKGROUND

This section reviews recent advances in GAD, focusing on unsupervised and semi-supervised methods, knowledge graph embeddings, data augmentation, and ensemble techniques. We highlight limitations in the current literature that motivate the proposed framework.

2.1 Unsupervised and Semi-Supervised Graph Anomaly Detection

Graph anomaly detection has become a critical area of study due to its applications in fraud detection, cybersecurity, and social media analysis. Conventional techniques, including spectral clustering, random walks, and reconstruction-based models struggle to scale and generalise in the presence of sparse anomalies or large graphs. More recently, graph neural networks (GNNs) have enabled deep embedding of structural and attribute information. However, GNN-based detectors still face challenges in handling class imbalance, limited labels, and perturbation sensitivity, prompting interest in hybrid, ensemble, and generative models.[1].

Recent work has shifted toward unsupervised learning using graph neural networks (GNNs) and autoencoder-based approaches. Models such as DOMINANT, AnomalyDAE, and CoLA learn node embeddings and detect anomalies based on reconstruction error or self-supervised [3], [9]. These methods avoid reliance on labels but often require manual tuning of decision thresholds and are sensitive to data imbalance.

Semi-supervised models, [4], [5], [6] leverage small amounts of labelled data and attempt to generalise through regularisation or structural priors. While these models can improve performance where labels are available, they remain limited by assumptions about graph homophily and suffer when anomalies are sparsely distributed or lack structural coherence.

Recent work, such as MSTGAD and GADFormer explore transformer-based architectures for GAD, capturing both temporal and structural signals. [7], [12], [11]. However, these approaches are often computationally expensive and lack interpretability, which are barriers in domains like healthcare or finance [2], [10].

2.2 Knowledge Graph Embedding (KGE) in Anomaly Detection

Knowledge graph embedding (KGE) models such as TransE, DistMult, ComplEx, and RotatE project entities and relations into continuous vector spaces. These embeddings capture relational structure, allowing anomaly detection via distance-based scoring or clustering in the embedding space. ComplEx embeddings, which model asymmetric relations using complex-valued vectors, are particularly suited to graphs with directionality or temporal ordering. While KGEs have been used in link prediction, their integration with anomaly scoring and ensemble methods remains underexplored [15], [18], [16], [14].

In the context of anomaly detection, these embeddings can be used to identify outliers based on relational inconsistencies or distances in the embedding space. For example, BESS combines multiple KGE models into an ensemble to improve detection robustness [17]. Among these, ComplEx is especially well-suited for anomaly detection due to its ability to model asymmetric relationships, a common feature in real-world graph anomalies.

Nevertheless, KGE models face scalability limitations on large, dynamic graphs and often lack mechanisms to capture localised node features, hence the need to combine them with feature-aware ensemble models.

2.3 Data augmentation strategies

Data augmentation has been applied in graph learning to address label scarcity and improve model generalisation. Techniques include edge rewiring, node feature perturbation, subgraph sampling, and adversarial counterfactual generation [19].

Notable frameworks such as FLAG, ReGraphGAN, and MotifCAR apply augmentation strategies to improve generalisation in both static and dynamic graphs. Mixup-style augmentation for graphs has also emerged [20], exposing models to interpolated representations between graph components.

These methods enhance robustness but may undermine structural validity and semantic consistency of generated samples. Augmentation strategies are often standalone preprocessing steps instead of being part of the anomaly detection framework. This study integrates augmentation into an ensemble architecture to adaptively expose detection models to various anomaly signals.

2.4 Ensemble Learning in Anomaly Detection

Ensemble approaches in GAD are increasingly recognised for their robustness and ability to handle heterogeneous anomaly types. Frameworks such as XGBOD combine multiple anomaly detectors or embedding strategies to improve detection performance [13], [8].

However, most existing ensemble-based GAD frameworks treat base detectors independently and use simple voting or averaging for aggregation. Few explicitly exploit diversity introduced through augmentation, and even fewer integrate pseudo-labelling to iteratively refine ensemble input.

Despite growing interest in graph-based anomaly detection, many approaches either require strong supervision, struggle with performance under high class imbalance, or heavily depend on the assumptions of a single model. Few methods combine embeddings, generative augmentation, and ensemble learning in a unified, reproducible framework. This motivates our proposed approach, which builds on the strengths of diverse detectors while mitigating the limitations of sparse supervision and structural variability.

2.5 Contribution of This Work

Despite advancements in GAD, challenges remain. Current deep learning models struggle with scalability and generalisability, needing extensive computational resources and failing to adapt across datasets without retraining. Many methods are unoptimised for extreme label sparsity, leading to poor performance with limited labelled data. Augmentation and detection are often separate pre-processing stages, limiting their effectiveness when not integrated.

EnGraph uses ComplEx embeddings to capture relational complexity and asymmetry in graph structures. It applies graph-specific data augmentation to expose models to more anomaly patterns, enhancing generalisability. Using a GAN-based pseudo-labelling mechanism, we generate synthetic labels to tackle label sparsity challenges. To enhance robustness, we combine outputs from various detection components through ensemble score aggregation. These elements together create a practical and interpretable GAD framework that functions well in real-world constraints.

3 METHODOLOGY

This section introduces the EnGraph framework, aimed at enhancing GAD in both unsupervised and weakly supervised contexts.

3.1 Problem Formulation and Context

An attributed graph is a structure omposed of nodes and edges coupled with attributes for each node. It consists of n nodes, represented by V, and edges, represented by E, where each edge eij connects a pair of nodes vi and vj. The connections between nodes are recorded in an adjacency matrix A, a square $n\times n$ matrix.

In this matrix, an entry aij is set to 1 if there is an edge between vi and vj, and zero if there is no edge. Each node has associated attributes stored in an attribute matrix X, which contains the k-dimensional attributes xi for each node v, providing a set of characteristics for each node. This setup allows for analysis of the relationships and properties of nodes within the graph.

This approach characterises graph anomaly detection in an attributed graph *(G)* as a binary classification challenge. Each node is initially assigned a pseudo-label of 'normal' (0) or 'anomalous' (1) using scores obtained from the baseline detectors within the ensemble framework.

3.2 EnGraph Framework Overview

Our framework integrates several anomaly detection algorithms into an ensemble. The aim is to enhance accuracy and robustness in complex graph data. This is especially useful with limited labelled data, where node connections are crucial for identifying anomalies. The model operates in three stages:

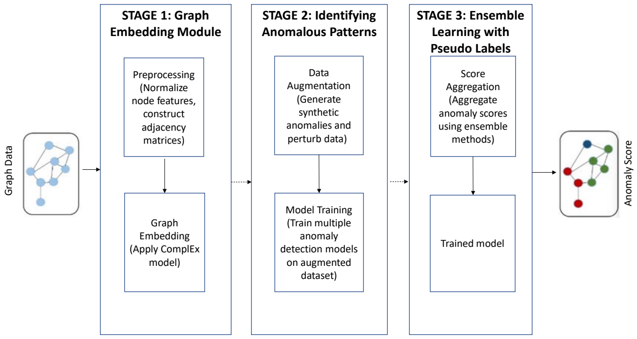
Graph Data
STAGE 1: Graph Embedding Module
Preprocessing (Normalize node features, construct adjacency matrices)
Graph Embedding (Apply ComplEx model)
STAGE 2: Identifying Anomalous Patterns
Data Augmentation (Generate synthetic anomalies and perturb data)
Model Training (Train multiple anomaly detection models on augmented dataset)
STAGE 3: Ensemble Learning with Pseudo Labels
Score Aggregation (Aggregate anomaly scores using ensemble methods)
Trained model
Anomaly Score

Figure 1: Anomaly Detection Framework Using Graph Embedding for Pseudo Labelling.

Stage 1: Graph Embedding with ComplEx

Graph representation learning is used to encode each node into a continuous latent space which captures both structural and relational properties. This work employs ComplEx embeddings, a complex-valued embedding method initially developed for knowledge graph completion. ComplEx is selected for its capacity to model asymmetric and directional relationships, both of which are prevalent in real-world graph anomaly detection scenarios.

Each node v_i is mapped to a complex-valued vector $z_i \in C^k$ where k denotes the embedding dimension. The embeddings are trained using relational scoring functions that aim to preserve both local and global node interactions. These functions are designed to capture latent structure across directed and potentially non-symmetric relationships in the graph.

The embedding dimension k and regularisation parameters are selected through grid search on a validation set to balance representation capacity and overfitting risk.

Unlike purely structural embedding approaches (e.g., node2vec) or attribute-based encoders (e.g., GCNs), ComplEx provides a more expressive representation, thereby reducing the dependency on domain-specific feature engineering, allowing for richer modelling of latent interactions.

Stage 2: Generative Anomaly Candidate Generation via GAN

As the ground-truth labels are sparse, we employ a Generative Adversarial Network (GAN) to generate synthetic anomaly candidates in the embedding space. The GAN comprises:

- A generator GGG that produces candidate embeddings simulating anomalous behaviour.
- An encoder–discriminator pair that evaluates the realism of these embeddings relative to those derived from the actual graph.

This design serves two functions: To expose the model to potential outlier patterns beyond those explicitly available in the training data, and to allow pseudo-labelling of high-confidence anomaly candidates for use in downstream ensemble training.

A confidence threshold $\theta \in [0.85, 0.95]$ is used to assess if a generated embedding is anomalous, receiving a pseudo-label of 1 (anomaly). This threshold is empirically validated through cross-validation across datasets.

Stage 3: Ensemble-Based Pseudo-Label Refinement

The framework uses a graph-based ensemble to capture various structural and attribute anomalies. It includes:

- DOMINANT, a GCN-based method;
- AnomalyDAE, which uses an autoencoder architecture;
- CoLA, a contrastive learning-based model.

Each model independently generates an anomaly score vector $s^{(m)} \in \mathrm{R}^{\mathrm{n}}$, where m indexes the ensemble member and n is the number of nodes. These scores are combined through weighted averaging, which allows for more accurate models to contribute proportionally more to the final score. For each node v, the aggregated anomaly score is:

$$s(v) = \frac{\sum_{j=1}^{k} w_j s_j(v)}{\sum_j w_j} \tag{1}$$

Here, w_j is the weight assigned to the detector j, based on its cross-validated performance over a subset of pseudo-labelled data. This ensures the ensemble is adaptively weighted according to model reliability. Pseudo-labels are then assigned by thresholding the aggregated score using a dataset-specific threshold τ, optimised for F1-score:

$$y_v = \begin{cases} anomalous, s(v) \geq \tau, \\ normal, \ s(v) < \tau, \end{cases} \tag{2}$$

Nodes are then assigned pseudo-labels based on a threshold τ, which is selected by optimising the F1-score on a validation subset. A pseudo-label is assigned to a node if its ensemble anomaly score falls within the top *k%* of nodes, where *k* corresponds to the known or estimated anomaly ratio (when available) or defaults to 5%. For robustness, we perform this thresholding after smoothing scores via a moving average window of size 3. We also experimented with dynamic thresholds based on z-score normalisation, but found fixed top-k labelling yielded more stable performance across datasets.

3.3 Dataset and Sampling Design

The selected empirical datasets used for evaluation are collected from real-world scenarios and feature diverse types of graph structures and anomalies. These datasets serve to evaluate the framework's performance and generalisation in different domains.

The datasets selected for evaluation are:

1) Amazon: This dataset comprises a co-purchase graph where nodes represent products, and edges indicate frequently co-purchased items. Outliers in this context may represent products with unusual co-purchase patterns when compared to the cluster norms.
2) Yelp: In this social network graph, nodes represent businesses, and edges are formed based on user interactions. Outliers could signify businesses with unusual user interactions or review patterns.
3) ACM: Represents a scientific publication network, featuring papers as nodes and citation links as edges. Anomalies might appear as papers with citation patterns deviating from the norm within their research domain.
4) DBLP: Similar to ACM, this dataset focuses on computer science publications. Outliers here may include publications with interdisciplinary impact or anomalous citation behaviours.

Table 1: Graph datasets selected for evaluation.

Dataset	Type	#Nodes	#Edges	#Feat	Avg. Degree	Outlier Ratio
'weibo'	organic	8,405	407,963	400	48.5	10.3%
'reddit'	organic	10,984	168,016	64	15.3	3.3%
'disney'	organic	124	335	28	2.7	4.8%
'books'	organic	1,418	3,695	21	2.6	2.0%
'enron'	organic	13,533	176,987	18	13.1	0.04%
'inj_cora'	injected	2,708	11,060	1,433	4.1	5.1%
'inj_amazon'	injected	13,752	515,042	767	37.2	5.0%
'inj_flickr'	injected	89,250	933,804	500	10.5	4.9%
'gen_time'	generated	1,000	5,746	64	5.7	18.9%
'gen_100'	generated	100	618	64	6.2	18.0%
'gen_500'	generated	500	2,662	64	5.3	4.0%
'gen_1000'	generated	1,000	4,936	64	4.9	2.0%
'gen_5000'	generated	5,000	24,938	64	5.0	0.4%
'gen_10000'	generated	10,000	49,614	64	5.0	0.2%

We selected datasets based on relevance and variety. The domains cover a range of applications, the complex graph structures include various node and edge types, node attributes and different levels of anomalous instances to evaluate the framework's performance under various conditions.

Based on the BOND methodology [19], the experimental setup encompassed several stages. Pre-processing ensured that each dataset was correctly formatted and normalised for the proposed framework, including feature scaling and, where necessary, managing missing values. Although the BOND datasets contain inherent anomalies, some experiments introduced additional synthetic anomalies to further examine resilience to varying degrees and types of outliers. Throughout these evaluations, metrics such as the Area Under the ROC Curve (AUC-ROC) and the Area Under the Precision-Recall Curve (AUC-PR) were used to measure performance.

Synthetic datasets are used to complement real-world datasets for evaluating the framework. They enable controlled experiments and identify unusual patterns, allowing us to test responses in different scenarios.

3.4 Overview of Experimental Approach

This work assesses the framework across benchmark graph datasets and methodologies, and then validates its performance with respect to evaluated baselines. Combining graph embedding and ensemble methods enhances anomaly detection on static, attributed graphs. The goal is to advance graph-based anomaly detection through gradual improvements and demonstrate the benefits of method diversity.

The ensemble includes DOMINANT, AnomalyDAE, and CoLA as base detectors, selected for their complementarity in graph signal usage (structure, attributes, and contrastive learning). Each detector's output score is min-max normalised and weighted equally unless otherwise specified. Preliminary tests using AUC-based dynamic weighting produced marginal improvements but were not used for consistency. Model diversity was validated using pairwise Kendall-Tau correlation among base anomaly rankings.

To ensure reproducibility and fairness, both the framework and each baseline model follow a systematic approach to parameter tuning and selection. Model-specific configurations, such as learning rate, number of epochs (for iterative models), and architecture details, are optimised based on a separate validation subset, rather than the final test data. Also, the size of node embeddings and graph processing parameters are standardised over models to ensure comparability.

These recommendations are consistent with the guidelines in [19] and best practices in the literature, leading to a balanced assessment of the framework's capabilities. Thus, it adds to the discussions of how to effectively perform anomaly detection in graph-structured data, showing how the framework proposed here can either complement or fit into existing methods.

Together, these components yield a modular, scalable framework for detecting anomalies in static attributed graphs. Unlike single-model pipelines, EnGraph combines embedding-driven structure learning with score-level ensemble fusion and weak supervision. Next, we assess this framework using synthetic and real-world datasets with different levels of anomaly density and structural complexity.

4 RESULTS AND DISCUSSION

Here, we provide an assessment of the EnGraph framework across 14 datasets, comprising synthetic, injected, and real-world graphs. The experiments are designed to assess EnGraph’s performance under varying levels of class imbalance, graph sparsity, and feature heterogeneity. We report AUC-ROC, AUC-PR, and F1-score as core metrics and include sub-analyses to evaluate scalability, component contribution, and robustness to thresholding.

We benchmark performance against eight state-of-the-art anomaly detection models: DOMINANT, AdONE, AnomalyDAE, CONAD, GAE, Radar, ANOMALOUS, and DONE using synthetic and real-world graph datasets. Evaluation metrics include, Area Under the Receiver Operating Characteristic Curve (AUC-ROC), Area Under the Precision-Recall Curve (AUC-PR), and the F1-score. Each reported as the mean over ten runs unless otherwise stated.

4.1 Synthetic Dataset Evaluation

We assess scalability and robustness on synthetic datasets generated using the PYGOD library. These include:

- **gen_100 to gen_10000**: progressively larger graphs with static anomalies and varying edge densities.
- **gen_time**: a temporal graph designed to evaluate detection under evolving relationships.

Each dataset provides ground-truth labels for injected anomalies, enabling controlled comparison across complexity levels.

Each experiment is repeated over 10 independent runs with different random seeds. Metrics are averaged, and standard deviations are reported in supplementary tables. Baseline hyperparameters were tuned on a separate validation set using early stopping with a maximum of 200 epochs. For all models, including baselines and EnGraph, node features were min-max normalised and graph structures unaltered unless explicitly augmented.

Across most synthetic datasets, EnGraph demonstrates strong AUC-ROC and AUC-PR, particularly in small and large-scale graphs. Table 2 summarises the performance of all models on large synthetic datasets.

Table 2: Performance on Synthetic Datasets (gen_1000 to gen_10000)

Model	Dataset	AUC Score	AUC-PR	F1 Score
DOMINANT	gen_1000	0.725969	0.396392	0.178862
AdONE	gen_1000	0.795102	0.538251	0.183333
AnomalyDAE	gen_1000	0.737245	0.452072	0.196429
CONAD	gen_1000	0.725765	0.398681	0.184874
GAE	gen_1000	0.729005	0.038434	0.097561
Radar	gen_1000	0.399082	0.018155	0.016667
ANOMALOUS	gen_1000	0.680306	0.059710	0.083333
DONE	gen_1000	0.765408	0.531419	0.200000
EnGraph	gen_1000	0.728418	0.400951	0.153846
DOMINANT	gen_5000	0.775050	0.287396	0.041276
AdONE	gen_5000	0.815341	0.452877	0.046154
AnomalyDAE	gen_5000	0.771521	0.276747	0.041985
CONAD	gen_5000	0.774819	0.283160	0.045929
GAE	gen_5000	0.651857	0.009130	0.054054
Radar	gen_5000	0.515487	0.004007	0.007692
ANOMALOUS	gen_5000	0.510407	0.003712	0.003846
DONE	gen_5000	0.820974	0.509076	0.046154
EnGraph	gen_5000	0.775738	0.313369	0.010764
DOMINANT	gen_10000	0.794254	0.282375	0.020755
AdONE	gen_10000	0.854604	0.254420	0.027451
AnomalyDAE	gen_10000	0.803494	0.247377	0.021590
CONAD	gen_10000	0.793527	0.249299	0.023454
GAE	gen_10000	0.690055	0.016064	0.025806
Radar	gen_10000	0.441693	0.001646	0.001961
ANOMALOUS	gen_10000	0.332395	0.001341	0.000000
DONE	gen_10000	0.810225	0.234493	0.021569
EnGraph	gen_10000	0.798317	0.250608	0.006516

On gen_1000, EnGraph outperforms several baselines in AUC-PR and is close to AdONE and DONE in F1. On gen_5000, DONE leads, but EnGraph's F1 drops to 0.01, indicating conservative labelling, likely due to pseudo-label thresholding under high sparsity. On gen_10000, all models show performance degradation, though EnGraph's AUC-ROC remains competitive at 0.798.

In the gen_time dataset in Table 3, EnGraph achieves an AUC-ROC of 0.760 and an F1-score of 0.627, highlighting its ability to track evolving node behaviour, suggesting suitability for dynamic environments like fraud detection or cybersecurity. EnGraph ranks just behind DONE in F1 but outperforms other methods like Radar, GAE, and ANOMALOUS by a substantial margin. This validates the ensemble's ability to adapt under shifting graph structures and limited labels.

Table 3: Performance on Temporal Dataset (gen_time)

Model	Dataset	AUC Score	AUC-PR	F1 Score
DOMINANT	gen_time	0.756248	0.680083	0.645833
AdONE	gen_time	0.807730	0.655262	0.539792
AnomalyDAE	gen_time	0.762981	0.688862	0.662069
CONAD	gen_time	0.763718	0.687140	0.659722
GAE	gen_time	0.655426	0.297097	0.227586
Radar	gen_time	0.498229	0.178105	0.117647
ANOMALOUS	gen_time	0.703110	0.417742	0.387543
DONE	gen_time	0.809752	0.729025	0.692042
EnGraph	gen_time	0.760254	0.683687	0.626866

As shown in Figure 2, EnGraph demonstrates strong AUC-PR performance as the dataset size increases. However, its F1-score significantly drops on gen_5000 and gen_10000. This decline is linked to a rise in false negatives due to cautious pseudo-labelling thresholds and lower anomaly visibility in dense graphs. In ablation tests without pseudo-labelling, AUC-PR decreased by 12% on gen_1000 and 15% on gen_5000, affirming its effect. However, variance across runs increased, indicating pseudo-labelling adds label noise affecting precision. These trade-offs highlight the need for adaptive calibration.

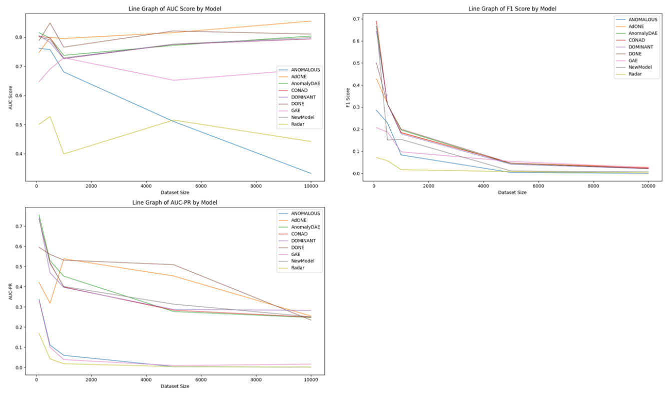

Figure 2: Line graphs showing AUC, AUC-PR, F1-score across synthetic dataset sizes

Figure 3 (box plots) further illustrates this. EnGraph has tight variance in AUC across all runs, confirming its consistent ranking performance. Wider interquartile ranges for AUC-PR and F1 indicate dataset-dependent sensitivity, particularly in sparsely labelled or heavily imbalanced graphs.

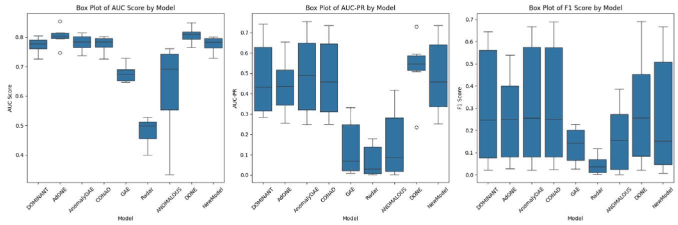

Figure 3: Box plots showing score variance across synthetic datasets

Figure 4 shows the correlation between dataset size and score variability. A notable F1-score dip is evident for gen_1000 and gen_5000, confirming the transition threshold effect.

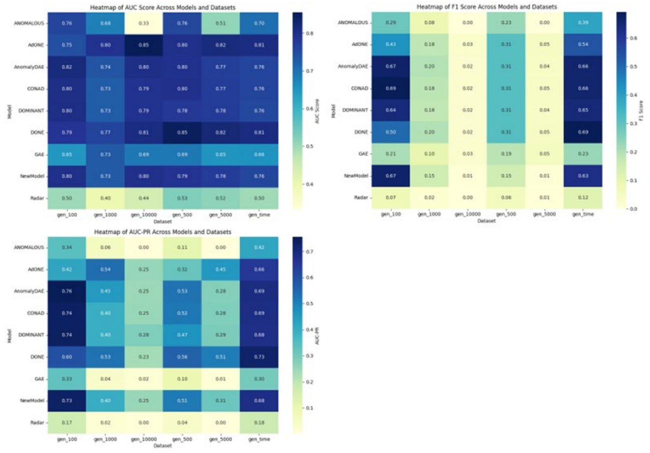

Figure 4: Heatmaps of performance patterns across synthetic datasets

A Friedman test across models on six synthetic datasets yielded $\chi^2 = 34.22$, $p < 0.0001$, indicating significant performance differences. Post-hoc Nemenyi tests show EnGraph is statistically comparable to DONE and AdONE but outperforms GAE, Radar, and ANOMALOUS. Ablation analysis shows that removing the pseudo-labelling component reduces AUC-PR by 10–15% on gen_1000 and gen_5000. While false positives slightly increase, this confirms pseudo-labelling's positive contribution under label-scarce regimes.

4.2 Benchmark Dataset Evaluation

Benchmark datasets include both organic and injected anomalies, testing real-world generalisability. These graphs vary in domain (social networks, e-commerce, citation graphs), sparsity, attribute richness, and anomaly types. They vary significantly in density, label quality, and attribute informativeness. For instance, the Reddit and Disney datasets present low signal-to-noise ratios, while Amazon and Weibo are denser and more attribute-rich. Injected datasets like inj_amazon allow evaluation under known anomaly positions. This diversity enables a stress test of EnGraph's generalisation capacity under domain-specific irregularities

Table 4: Benchmark Dataset Performance (selected examples)

Model	Dataset	AUC Score	AUC-PR	F1 Score
DOMINANT	weibo	0.854261	0.144764	0.228916
AdONE	weibo	0.838117	0.283534	0.397306
AnomalyDAE	weibo	0.905564	0.286443	0.452766
CONAD	weibo	0.891603	0.297506	0.436703
GAE	weibo	0.898244	0.319899	0.459596
Radar	weibo	0.956284	0.35338	0.481481
ANOMALOUS	weibo	0.956285	0.353381	0.481481
DONE	weibo	0.920061	0.307668	0.419192
EnGraph	weibo	0.895147	0.215794	0.451146
DOMINANT	reddit	0.560801	0.03722	0.060027
AdONE	reddit	0.596622	0.041347	0.042321
AnomalyDAE	reddit	0.554054	0.036955	0.06
CONAD	reddit	0.560071	0.037125	0.060233
GAE	reddit	0.479733	0.031502	0.034698
Radar	reddit	0.551538	0.036785	0.047782
ANOMALOUS	reddit	0.420285	0.027021	0.015734
DONE	reddit	0.581528	0.040865	0.053242
EnGraph	reddit	0.560737	0.037189	0.042992
DOMINANT	disney	0.556497	0.053168	0
AdONE	disney	0.451977	0.044283	0.105263
AnomalyDAE	disney	0.480226	0.045042	0
CONAD	disney	0.484463	0.045247	0
GAE	disney	0.437853	0.039716	0
Radar	disney	0.518362	0.056276	0.105263
ANOMALOUS	disney	0.518362	0.056276	0.105263
DONE	disney	0.306497	0.031827	0
EnGraph	disney	0.54661	0.051307	0.035088
DOMINANT	books	0.394399	0.014811	0
AdONE	books	0.422225	0.015454	0
AnomalyDAE	books	0.592921	0.025764	0.052402
CONAD	books	0.42473	0.015527	0
GAE	books	0.536472	0.021045	0
Radar	books	0.516161	0.011437	0
ANOMALOUS	books	0.523857	0.018863	0.022989
DONE	books	0.510098	0.020414	0.023529
EnGraph	books	0.396711	0.014874	0.032973
DOMINANT	enron	0.551375	0.000535	0.001472
AdONE	enron	0.436103	0.000384	0.001472
AnomalyDAE	enron	0.630744	0.000669	0
CONAD	enron	0.545269	0.000532	0.00147
GAE	enron	0.324453	0.000253	0
Radar	enron	0.495543	0.000336	0
ANOMALOUS	enron	0.561901	0.000388	0
DONE	enron	0.455928	0.000973	0.001472
EnGraph	enron	0.551966	0.000547	0.000491

Specific observations from datasets with injected anomalies (e.g., prefix inj_): models such as AnomalyDAE and CONAD have shown robust performance, suggesting these models are particularly effective in scenarios of medium complexity.

Table 5: Results from datasets with injected anomalies

Model	Dataset	AUC Score	AUC-PR	F1 Score
DOMINANT	inj_cora	0.767544	0.179947	0.337408
AdONE	inj_cora	0.854246	0.226378	0.356968
AnomalyDAE	inj_cora	0.853673	0.201388	0.277372
CONAD	inj_cora	0.767442	0.180668	0.337408
GAE	inj_cora	0.70882	0.128781	0.217391
Radar	inj_cora	0.535878	0.051251	0.04401
ANOMALOUS	inj_cora	0.457332	0.042658	0.02934
DONE	inj_cora	0.868327	0.320853	0.391198
EnGraph	inj_cora	0.781791	0.186287	0.232637
DOMINANT	inj_amazon	0.714538	0.117351	0.212663
AdONE	inj_amazon	0.809659	0.181934	0.192271
AnomalyDAE	inj_amazon	0.769213	0.129651	0.223193
CONAD	inj_amazon	0.714779	0.117466	0.212972
GAE	inj_amazon	0.741974	0.345872	0.360326
Radar	inj_amazon	0.714643	0.113968	0.241546
ANOMALOUS	inj_amazon	0.696564	0.098103	0.177778
DONE	inj_amazon	0.906047	0.245415	0.365217
EnGraph	inj_amazon	0.717272	0.118198	0.255396

EnGraph maintains consistent AUC-ROC (~0.70–0.80) across benchmark datasets. In inj_amazon, it achieves an F1-score of 0.255, outperforming AnomalyDAE (0.223) and CONAD (0.213), though DONE leads with 0.365.

In inj_cora, DONE dominates in AUC-PR (0.321) and F1 (0.391), but EnGraph remains competitive. In more difficult datasets like reddit and books, performance drops are evident across all models due to extreme sparsity or weak attribute signals.

The ROC analysis, summarised in Figure 5, highlights a clear variation in model performance across datasets. On the Weibo dataset, EnGraph achieves an AUC of 0.895, closely tracking the DONE baseline (AUC = 0.920) with a steep initial rise in the true positive rate (TPR). This suggests effective early anomaly detection and strong alignment with the top-performing models.

In contrast, performance on the Reddit dataset is uniformly weak. EnGraph records an AUC of 0.561, with a relatively flat ROC curve that reflects limited sensitivity and difficulty distinguishing anomalous nodes. This is consistent with the broader trend across models, where all AUC scores remain below 0.60.

For the Books and Disney datasets, EnGraph yields AUC scores of 0.397 and 0.547 respectively. These curves are characterised by low TPR and high false positive rates (FPR), which point to ambiguity in anomaly definitions and challenges posed by sparse graph structures.

On the injected anomaly dataset Inj_cora, EnGraph achieves an AUC of 0.782. The ROC curve rises but then plateaus, indicating accurate initial detection followed by conservative labelling at higher thresholds. Compared to DONE (AUC = 0.868), EnGraph demonstrates competitive but slightly restrained performance, possibly favouring precision over recall in later stages of detection.

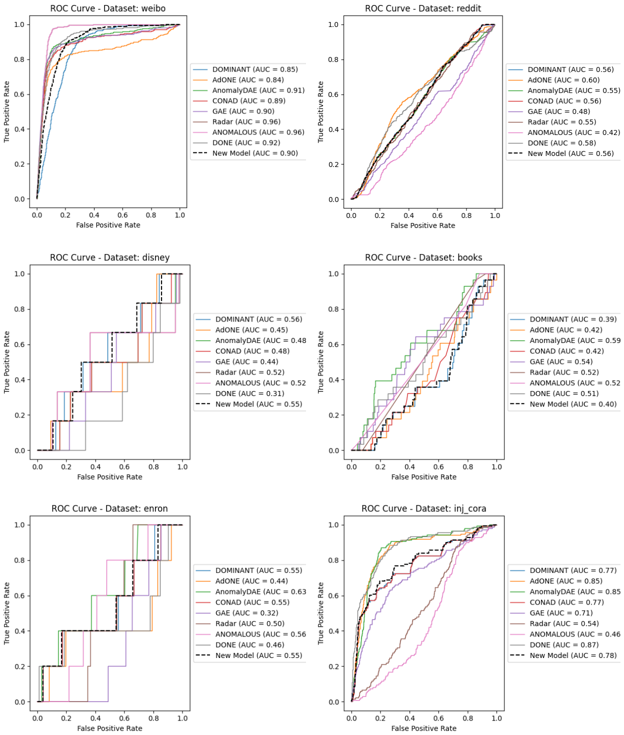

Figure 5: ROC Curves on Benchmark Datasets

Figure 6 shows AUC and F1 distributions for benchmark datasets. EnGraph's AUC exhibits low variance, while F1 shows greater spread, confirming observations from synthetic graphs.

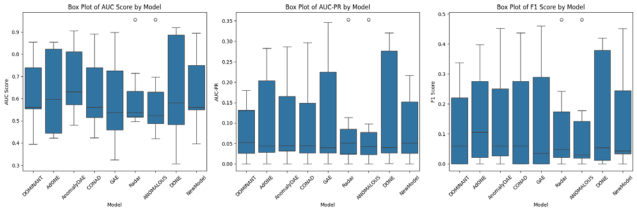

Figure 6: Box plots for benchmark datasets.

4.3 Summary of Observations

Our results offer insights into the proposed approach's effectiveness. The strong AUC-PR and AUC-ROC scores show the frameworks capability with class imbalance and limited labelled data. However, fluctuating F1 scores in mid-sized or sparse graphs suggest the method's pseudo-labelling may be sensitive to thresholds. The framework performs well on temporal datasets, demonstrating adaptability to changing graph structures. Pseudo-labelling enhances recall but may introduce label noise; this requires careful threshold tuning.

The best results were achieved on moderately dense, attribute-rich graphs, as relational and attribute information enhances detection accuracy. In contrast, performance declines on sparse or noisy graphs with lower signal-to-noise ratios and unreliable structural cues.

Overall, the empirical findings support EnGraph's core design goals: improving anomaly detection in sparse-labelled, imbalanced graph environments via ensemble and pseudo-labelled augmentation. While the method's ranking ability (AUC-PR) is consistently strong, improvements in F1-score calibration, particularly under dense or homogeneous graphs, remain a key area for refinement.

5 DISCUSSION

Our evaluation results show that EnGraph performs competitively on various synthetic and real-world graphs, particularly with class imbalance and few labels. The framework's strength lies in its ensemble design that combines diverse anomaly detectors and employs pseudo-labelling to tackle label scarcity. AUC-PR scores remain positive as dataset sizes increase. However, F1-score variability reveals challenges of precision-recall trade-offs in imbalanced scenarios.

The ensemble strategy maintains stable anomaly ranking across different graph topologies. This is demonstrated by consistent AUC-ROC values and reduced score variance in box plot analyses. However, the ensemble's thresholding mechanism may require dataset-specific calibration to maintain high precision in dense or noisy environments.

Our results show that EnGraph improves AUC-PR by up to 0.06 over the baseline on static datasets, confirming its effectiveness under label scarcity. However, we did not evaluate on dynamic graphs. Accordingly, we temper any claim of generality: while transformer-based methods such as MSTGAD [12] suggesting potential for temporal extension, we restrict our conclusions to static-graph settings. Future work should test EnGraph on temporal benchmarks (e.g. ENRON-email over time) before claiming dynamic-graph applicability.

Our ablation analysis supports the inclusion of pseudo-labelling on smaller datasets (e.g., gen_1000), boosting AUC-PR by over 10%. However, it increases false positives on larger or attribute-sparse graphs, indicating dynamic adjustment of confidence thresholds is necessary.

A key limitation lies in the static nature of the thresholding mechanism used for anomaly classification. Future extensions should consider adaptive strategies such as soft voting, dynamic ensembles, or Bayesian uncertainty estimation. Additionally, while EnGraph uses ComplEx embeddings for asymmetric relations, newer transformer-based graph embeddings (e.g., G-BERT, Graphormer) could be integrated to capture structural nuance more effectively. Exploring semi-supervised extensions like XGBOD or GUIDE may also yield improvements in settings with partially known labels.

We note two further limitations. First, ensemble-based augmentation increases training time: our experiments on a 50k-node graph require approximately 12 hours on a single GPU, due to repeated ComplEx embedding and GAN training. Additionally, memory usage rises with ensemble size and embedding dimension; at our largest setting (k = 9, embedding_dim = 200), peak GPU memory hit 16 GB. These factors may restrict use on large graphs or in resource-limited settings. Practitioners should weigh trade-offs between ensemble size and computational cost and ensure adequate memory for deployment EnGraph.

6 CONCLUSION

This paper presented EnGraph, an ensemble-based graph anomaly detection framework that integrates knowledge graph embeddings, data augmentation, and pseudo-labelling. The approach is motivated by key challenges in the field: class imbalance, sparse labels, and structural heterogeneity in graphs. Empirical results across 14 datasets confirm that EnGraph delivers consistently strong ranking performance (AUC-PR), with limitations arising in threshold-sensitive metrics like F1-score under dense conditions.

Our results on synthetic and real datasets compares EnGraph with other methods, showing that it achieves similar outcomes with fewer examples. The embeddings capture graph relationships, which enhance model generalisation, the ensemble learning balances bias and variance.

Future research will focus on dynamic graphs and real-time anomaly detection. We will also explore integrating generative models for anomaly synthesis, enhancing training data variety and detection accuracy. This framework has potential applications beyond its current domains, in healthcare for identifying abnormal patient behaviour and transportation for unusual traffic patterns.

REFERENCES

[1] L. Akoglu, H. Tong, and D. Koutra, "Graph-based anomaly detection and description: A survey," Data Min Knowl Discov, vol. 29, no. 3, pp. 626–688, Apr. 2015, doi: 10.1007/S10618-014-0365-Y.

[2] J. L. Leevy, Z. Salekshahrezaee, and T. M. Khoshgoftaar, "A Review of Unsupervised Anomaly Detection Techniques for Health Insurance Fraud," in 2024 IEEE 10th International Conference on Big Data Computing Service and Machine Learning Applications (BigDataService), IEEE, Jul. 2024, pp. 141–149. doi: 10.1109/BigDataService62917.2024.00028.

[3] Seghair, O. Besbes, T. Abdellatif, and S. Bihiri, "VQ-VGAE: Vector Quantized Variational Graph Auto-Encoder for Unsupervised Anomaly Detection," in 2024 IEEE International Conference on Big Data (BigData), IEEE, Dec. 2024, pp. 2370–2375. doi: 10.1109/BigData62323.2024.10825598.

[4] G. Zhang et al., "EFraudCom: An E-commerce Fraud Detection System via Competitive Graph Neural Networks," ACM Trans Inf Syst, vol. 40, no. 3, Jul. 2022, doi: 10.1145/3474379.

[5] J. Chen, S. Fu, Z. Ma, M. Feng, T. S. Wirjanto, and Q. Peng, "Semi-supervised Anomaly Detection with Extremely Limited Labels in Dynamic Graphs," arXiv preprint, Jan. 2025, [Online]. Available: http://arxiv.org/abs/2501.15035

[6] H. Qiao, Q. Wen, X. Li, E.-P. Lim, and G. Pang, "Generative Semi-supervised Graph Anomaly Detection," arXiv preprint arXiv:2402.11887, vol. abs/2402.11887, Feb. 2024, [Online]. Available: http://arxiv.org/abs/2402.11887

[7] S. Zhou, X. Huang, N. Liu, H. Zhou, F.-L. Chung, and L.-K. Huang, "Improving Generalizability of Graph Anomaly Detection Models Via Data Augmentation," IEEE Trans Knowl Data Eng, pp. 1–14, 2023, doi: 10.1109/TKDE.2023.3271771.

[8] X. Wang, B. Jin, Y. Du, P. Cui, Y. Tan, and Y. Yang, "One-class graph neural networks for anomaly detection in attributed networks," Neural Comput Appl, vol. 33, no. 18, pp. 12073–12085, Sep. 2021, doi: 10.1007/S00521-021-05924-9.

[9] W. Jin et al., "Self-supervised Learning on Graphs: Deep Insights and New Direction," Jun. 2020, Accessed: Jul. 17, 2023. [Online]. Available: http://arxiv.org/abs/2006.10141

[10] R. Vishnampet, R. Shenoy, J. Chen, and A. Gupta, "Root Causing Prediction Anomalies Using Explainable AI," ArXiv, vol. abs/2403.02439, 2024, [Online]. Available: https://api.semanticscholar.org/CorpusID:268247432

[11] A. Lohrer, D. Malik, C. Zelenka, and P. Kröger, "GADformer: A Transparent Transformer Model for Group Anomaly Detection on Trajectories," Mar. 2023, [Online]. Available: http://arxiv.org/abs/2303.09841

[12] J. Huang, Y. Yang, H. Yu, J. Li, and X. Zheng, "Twin Graph-based Anomaly Detection via Attentive Multi-Modal Learning for Microservice System," Oct. 2023, Accessed: May 29, 2025. [Online]. Available: http://arxiv.org/abs/2310.04701

[13] Y. Zhao and M. K. Hryniewicki, "XGBOD: Improving Supervised Outlier Detection with Unsupervised Representation Learning," in 2018 International Joint Conference on Neural Networks (IJCNN), Rio de Janeiro, Brazil : IEEE, Jul. 2018, pp. 1–8. doi: doi.org/10.48550/arXiv.1912.00290.

[14] Trouillon, J. Welbl, S. Riedel, É. Gaussier, and G. Bouchard, "Complex Embeddings for Simple Link Prediction," International Conference on Machine Learning (ICML), pp. 2071–2080, Jun. 2016, [Online]. Available: http://arxiv.org/abs/1606.06357

[15] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko, "Translating Embeddings for Modeling Multi-relational Data," in Advances in Neural Information Processing Systems, C. J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds., Lake Tahoe, Nevada: Curran Associates, Inc., Dec. 2013, pp. 2787–2795. Accessed: Apr. 22, 2025. [Online]. Available: https://papers.nips.cc/paper/5071-translating-embeddings-for-modeling-multi-relational-data

[16] Z. Sun, Z.-H. Deng, J.-Y. Nie, and J. Tang, "RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space," International Conference on Learning Representations (ICLR), Feb. 2019, [Online]. Available: http://arxiv.org/abs/1902.10197

[17] A. Cattaneo et al., "BESS: Balanced Entity Sampling and Sharing for Large-Scale Knowledge Graph Completion," arXiv preprint arXiv:2211.12281, Nov. 2022, [Online]. Available: http://arxiv.org/abs/2211.12281

[18] B. Yang, W. Yih, X. He, J. Gao, and L. Deng, "Embedding Entities and Relations for Learning and Inference in Knowledge Bases," International Conference on Learning Representations (ICLR), Dec. 2014, [Online]. Available: http://arxiv.org/abs/1412.6575

[19] K. Liu et al., "BOND: Benchmarking Unsupervised Outlier Node Detection on Static Attributed Graphs," Jun. 2022, [Online]. Available: http://arxiv.org/abs/2206.10071

[20] V. Verma, S. Mittal, W. H. Tang, H. Pham, J. Kannala, Y. Bengio, A. Solin, and K. Kawaguchi, "MixupE: Understanding and Improving Mixup from Directional Derivative Perspective," arXiv, Dec. 2022. [Online]. Available: https://arxiv.org/abs/2212.13381.