

A comparative study of GCN and GraphSAGE algorithms for structural anomaly detection in public procurement networks

Danny Mario Tua Panjaitan¹, Eka Prima Ginting²

^{1,2}Magister Teknik Informatika, Universitas Sumatera Utara, Indonesia

Email: dannypanjaitan0@gmail.com; ekaprima@students.usu.ac.id

ABSTRAK

Manipulasi tender dan kolusi pengadaan dapat menghasilkan pola struktural dalam hubungan antara perusahaan dan peluang pengadaan. Studi ini membandingkan *Graph Convolutional Network* (GCN) dan GraphSAGE untuk deteksi anomali struktural dalam jaringan pengadaan perusahaan-tender yang heterogen. Penelitian ini menggunakan data pengadaan pemerintah melalui GeBIZ dari Singapura dan memperkenalkan anomali sintetis terkontrol karena dataset tersebut tidak menyediakan informasi partisipasi peserta tender yang lengkap ataupun label kolusi yang terverifikasi. Evaluasi temporal yang mempertimbangkan risiko kebocoran data diterapkan bersamaan dengan variasi tingkat kesulitan anomali, *random seed*, subjek anomali yang belum pernah ditemui sebelumnya, serta pengujian lintas pola. Hasil penelitian menunjukkan adanya perbedaan perilaku empiris antara kedua arsitektur tersebut. GraphSAGE mencapai nilai F1-Score, ROC-AUC, dan PR-AUC yang lebih tinggi dalam evaluasi *multi-seed*, sedangkan GCN mencapai nilai *recall* yang lebih tinggi. GraphSAGE juga menunjukkan kinerja lebih baik dalam evaluasi induktif murni terhadap subjek anomali yang belum pernah ditemui sebelumnya. Sebuah Sistem Peringatan Dini berbasis GraphSAGE dikembangkan lebih lanjut untuk memeringkat perusahaan berdasarkan skor anomali struktural mereka. Temuan ini mengindikasikan bahwa model berbasis graf dapat mendukung penyaringan anomali struktural, namun hasil anomali sintetis tidak boleh ditafsirkan sebagai bukti adanya kolusi pengadaan yang terkonfirmasi.

Kata Kunci: pengaturan tender; jaringan saraf graf; pengadaan publik; deteksi anomali struktural; sistem peringatan dini GraphSAGE

ABSTRACT

Bid rigging and procurement collusion may produce structural patterns in the relationships between companies and procurement opportunities. This study compares Graph Convolutional Network (GCN) and GraphSAGE for structural anomaly detection in a heterogeneous company-tender procurement network. The study uses Government Procurement via GeBIZ data from Singapore and introduces controlled synthetic anomalies because the dataset does not provide complete bidder-participation information or verified collusion labels. A leakage-aware temporal evaluation is applied together with multiple anomaly difficulties, random seeds, unseen anomaly subjects, and cross-pattern testing. The results show different empirical behavior between the two architectures. GraphSAGE achieves higher F1-Score, ROC-AUC, and PR-AUC in the multi-seed evaluation, while GCN achieves higher recall. GraphSAGE also performs better in the true inductive evaluation on unseen anomaly subjects. An Early Warning System based on GraphSAGE is further developed to rank companies according to their structural anomaly scores. The findings indicate that graph-based models can support structural anomaly screening, while synthetic anomaly results should not be interpreted as evidence of confirmed procurement collusion.

Keywords: bid rigging; graph neural networks; public procurement; structural anomaly detection; graphsage early warning system

Corresponding Author:

Danny Mario Tua Panjaitan,
Universitas Sumatera Utara,
alan Dr. T. Mansur No. 9, Padang Bulan, Indonesia
Email: dannypanjaitan0@gmail.com



1. INTRODUCTION

Public procurement involves repeated interactions among government agencies, suppliers, and tender opportunities, creating a relational structure that cannot always be adequately represented through isolated transaction attributes. One important concern in this environment is bid rigging, in which firms coordinate their participation or bidding behavior in ways that may reduce effective competition. Such behavior can leave structural traces in procurement networks, particularly through repeated relationships among companies and tenders. Previous research has demonstrated that structural information can be useful for identifying potentially collusive networks in electronic procurement, including the detection of suspicious relationships among firms participating in multiple tenders (Baránek et al., 2021). Therefore, representing procurement records as a graph provides an alternative perspective for analyzing patterns that may not be apparent from individual tabular observations.

Graph Neural Networks (GNNs) provide a suitable framework for learning from both node attributes and graph connectivity. A Graph Convolutional Network (GCN), introduced by (Kipf & Welling, 2017), propagates information between neighboring nodes so that node representations incorporate both local features and graph structure. This characteristic makes GCN suitable for node-level classification when the graph used for learning is sufficiently represented during training. However, procurement environments are not static: new companies and tender opportunities may appear over time. This creates an important distinction between transductive learning, in which the graph structure is known during model training, and inductive learning, in which the learned representation function must also be applicable to previously unseen nodes.

GraphSAGE was proposed by (Hamilton et al., 2017) as an inductive framework that learns a function for generating node representations through neighborhood sampling and aggregation. Unlike approaches that learn representations specifically for the nodes observed during training, GraphSAGE can generate representations for previously unseen nodes when their features and neighborhood information are available. This property makes the comparison between GCN and GraphSAGE relevant to procurement monitoring systems in which the underlying network evolves over time. However, the theoretical inductive capability of GraphSAGE does not by itself establish better anomaly-detection performance; its effectiveness must be evaluated empirically under conditions that explicitly contain unseen entities.

A further methodological issue concerns the evaluation of structural anomaly detection. Synthetic anomalies can be useful when a dataset does not contain sufficiently labeled real-world cases, because they provide controlled ground-truth labels and allow researchers to vary the characteristics of anomalous patterns. Nevertheless, highly structured synthetic anomalies may also be easier to distinguish than naturally occurring anomalies. In addition, severe class imbalance and graph message passing can create misleadingly optimistic results when the temporal and structural relationship between training and test observations is not carefully controlled. Consequently, an evaluation framework for procurement anomaly detection should examine not only conventional classification performance, but also robustness across anomaly difficulties, temporal generalization, unseen entities, and patterns that differ from those used during training.

This study addresses these issues by modeling a public procurement dataset from Singapore's Government Electronic Business (GeBIZ) platform covering procurement observations from 2015 to 2021 as a heterogeneous company–tender graph. Company and tender entities are represented as different node types, while observed supplier–tender relationships are represented in the graph. Because the dataset does not provide complete bidder-participation information for constructing directly observed co-bidding relationships, structural anomaly patterns are introduced through controlled synthetic edge injection. Importantly, these synthetic relationships are treated as experimental ground truth rather than as confirmed evidence of real-world collusion. The experimental design therefore focuses on evaluating whether GNNs can distinguish predefined structural anomaly patterns under controlled and leakage-aware conditions.

The main objective of this research is to conduct a controlled empirical comparison between GCN and GraphSAGE for structural anomaly detection in public procurement networks. The evaluation considers three anomaly difficulties—hard, moderate, and easy—and uses multiple random seeds to assess robustness. A temporal data partition is employed to separate historical training observations from validation and future test observations. The study additionally evaluates true inductive performance using companies that were not observed during training and examines cross-pattern generalization by training on a cartel-type structural pattern and testing on a different temporal burst pattern. Computational efficiency is also evaluated through training and inference time, while an Early Warning System (EWS) calibration procedure is developed using validation data without using the test set for threshold selection.

Accordingly, the contributions of this study are fourfold. First, it establishes a leakage-aware temporal evaluation framework for GNN-based structural anomaly detection in procurement networks. Second, it evaluates model robustness across multiple synthetic anomaly difficulties and random seeds rather than relying on a single highly structured anomaly configuration. Third, it explicitly examines inductive generalization to previously unseen companies and cross-pattern generalization to a structural anomaly pattern not used during

training. Fourth, it translates the resulting anomaly scores into an EWS-oriented screening output while maintaining a clear distinction between structural anomaly detection and the substantive determination of collusion. Through this framework, the study provides an empirical basis for understanding the respective behavior of GCN and GraphSAGE under evolving, imbalanced, and structurally heterogeneous procurement-network conditions.

2. LITERATURE REVIEW

Graph Neural Networks (GNNs) have become an important approach for learning representations from graph-structured data because they can jointly exploit node attributes and relational information. Unlike conventional machine learning methods that generally treat observations as independent samples, GNNs propagate information through graph connections, allowing the representation of a node to be influenced by its local neighborhood. One of the foundational architectures is the Graph Convolutional Network (GCN), which applies graph convolution by aggregating information from neighboring nodes to learn representations for node-level prediction (Kipf & Welling, 2017). GCN has subsequently become one of the widely adopted architectures for graph-based classification and representation learning, particularly when relationships among entities provide important contextual information beyond individual node attributes.

GraphSAGE extends neighborhood-based graph representation learning by learning an aggregation function that generates node representations from local neighborhood information. Unlike conventional transductive approaches, GraphSAGE is designed to generate representations for previously unseen nodes by learning a general aggregation function rather than an embedding for every individual node (Hamilton et al., 2017). This distinction between transductive and inductive learning is particularly relevant to public procurement networks because procurement relationships may evolve as new companies participate in procurement activities and new tender opportunities become available. Therefore, comparing GCN and GraphSAGE provides an opportunity to examine whether their different representation-learning mechanisms affect structural anomaly detection when the model encounters previously unseen entities.

The relational characteristics of procurement data further motivate the use of graph representations capable of modeling different entities and relationships. Public procurement networks can contain different types of entities, such as companies and tenders, together with relationships describing their interactions. Relational Graph Convolutional Networks (R-GCN), for example, were proposed to extend graph convolution to multi-relational data by incorporating relation-specific transformations (Schlichtkrull et al., 2018). Although the present study does not employ R-GCN, this work provides an important foundation for understanding how graph neural networks can represent relational structures involving different types of relationships. More recent research has also applied heterogeneous graph neural networks to fraud detection in supply-chain finance, demonstrating the potential of heterogeneous representations for capturing complex interactions among different types of entities (Wu et al., 2024).

Graph anomaly detection focuses on identifying unusual nodes, edges, or subgraph structures by considering both graph attributes and structural relationships. Recent surveys have established graph anomaly detection as an important research area within graph machine learning and have categorized anomalies according to different graph structures and anomaly types (Ma et al., 2023). Similarly, research specifically examining GNN-based graph anomaly detection has shown that GNNs can exploit node attributes and graph structure to learn representations for identifying anomalous patterns. Such approaches have been studied for static and dynamic graphs as well as node-, edge-, subgraph-, and graph-level anomalies (Kim et al., 2022). These studies provide a theoretical basis for using GNNs to detect structural irregularities rather than relying exclusively on conventional attribute-based anomaly detection.

The use of graph structures for fraud detection has also received considerable attention. A systematic review of graph-based anomaly detection for fraud detection emphasized that relationships and interdependencies among entities can provide important information for identifying suspicious behavior that may not be apparent from isolated observations (Pourhabibi et al., 2020). More recent research has specifically examined the application of GNNs to fraud detection and identified a broad range of graph settings, including homogeneous and heterogeneous graphs as well as static, temporal, and dynamic networks (Motie & Raahemi, 2024). These findings are relevant to public procurement because companies and tenders form interconnected networks in which potentially suspicious behavior may emerge from patterns of relationships rather than from a single transaction or company attribute.

Another important consideration is the structural characteristics of fraud graphs. Recent research has shown that fraud networks may exhibit heterophily, meaning that connected nodes do not necessarily belong to the same class. This condition can challenge GNN architectures that implicitly benefit from homophilic neighborhood structures. In addition, fraud detection commonly involves severe class imbalance, which makes the identification of minority-class anomalies particularly challenging (Xu et al., 2024). These observations are

important for the present study because potentially anomalous companies represent only a small proportion of the procurement network. Therefore, model evaluation should not rely solely on accuracy, but should also consider precision, recall, F1-score, ROC-AUC, and PR-AUC.

The temporal dimension of graph data is also relevant to anomaly detection in procurement networks. In real-world procurement systems, relationships between companies and tenders change over time, meaning that a model trained on historical relationships may subsequently encounter new structural configurations. Research on anomaly detection in dynamic graphs has therefore investigated methods that explicitly consider temporal changes in graph structures (Liu et al., 2023). The temporal perspective is particularly relevant to an early warning system because evaluation using randomly mixed observations may allow information from future procurement activities to influence the learning process. A temporal evaluation framework can instead provide a more realistic assessment of how a model performs when applied to later procurement periods.

Research specifically related to procurement and bidding has also developed through several methodological perspectives. A scientometric analysis of construction bidding research identified a broad evolution of research topics and analytical approaches within the bidding domain, demonstrating the increasing diversity of methods applied to procurement-related problems (Rashidi et al., 2023). More directly related to collusion, Baránek et al. investigated collusive networks in electronic procurement and demonstrated that relationships among firms can provide useful signals for identifying potentially collusive structures (Baránek et al., 2021). These studies support the relevance of network-based analysis for procurement research, although their objectives and analytical approaches differ from the direct GCN and GraphSAGE comparison conducted in the present study.

Other studies have investigated suspicious procurement and fraud-related patterns using different computational approaches. Zeng proposed a residual graph neural network architecture for large-scale fraud detection, demonstrating the potential of specialized GNN architectures for learning fraud-related structural patterns (Zeng, 2021). Bejger examined parallel pricing patterns using unsupervised techniques such as Dynamic Time Warping and Matrix Profile and identified unusual temporal pricing patterns as potential indicators of cartel behavior (Bejger, 2024). These studies demonstrate that suspicious behavior can be identified through both graph structure and temporal or behavioral patterns, but they do not directly evaluate GCN and GraphSAGE under a common procurement-network experimental framework.

Privacy-preserving graph-based fraud detection has also emerged as an important research direction. Li and Walsh investigated a combination of federated learning and graph attention mechanisms for fraud detection, illustrating how graph-based learning can be combined with privacy-preserving approaches in transaction networks (Li & Walsh, 2024). Although privacy preservation is outside the primary scope of the present study, this line of research demonstrates the broader applicability of graph neural networks to fraud-related problems in complex relational environments.

Taken together, the existing literature establishes several important foundations for the present research. GCN provides a fundamental architecture for graph-based representation learning (Kipf & Welling, 2017), while GraphSAGE introduces an inductive mechanism for generating representations of previously unseen nodes (Hamilton et al., 2017). Relational and heterogeneous graph studies demonstrate the suitability of graph learning for networks containing multiple entity and relationship types (Schlichtkrull et al., 2018) (Wu et al., 2024). Research on graph anomaly detection and fraud detection further establishes the importance of combining structural and attribute information (Ma et al., 2023) (Pourhabibi et al., 2020). Meanwhile, procurement and bidding studies demonstrate the relevance of structural and behavioral analysis for identifying potentially suspicious procurement patterns (Baránek et al., 2021) (Rashidi et al., 2023).

However, an empirical gap remains in the direct comparison of GCN and GraphSAGE for structural anomaly detection in public procurement networks under a common and leakage-aware evaluation framework. Existing studies have generally focused on individual fraud-detection architectures, procurement collusion analysis, temporal pricing behavior, or broader graph anomaly detection rather than simultaneously evaluating transductive performance, inductive generalization, class imbalance, temporal separation, and cross-pattern anomaly generalization within the same procurement network. The present study addresses this gap by comparing GCN and GraphSAGE using a heterogeneous procurement network derived from Government Procurement via GeBIZ data. The evaluation incorporates temporal data separation, controlled structural anomaly scenarios, unseen anomaly subjects, multiple random seeds, and different anomaly patterns. In addition to classification performance, the study evaluates training and inference efficiency and examines the application of the resulting model as an early warning screening mechanism.

3. RESEARCH METHOD

To ensure a rigorous comparison, the research is conducted through the following systematic phases:

A. Dataset and preprocessing

This study uses the Government Procurement via GeBIZ dataset from Singapore, covering procurement observations from 2015 to 2021. The original dataset contains 32,756 supplier–tender records with information on tender identification, description, agency, award date, status, supplier, and awarded amount. Records with an Unknown supplier were removed because they could not be reliably associated with an identifiable company. After preprocessing, 31,342 supplier–tender relationships involving 8,632 identifiable companies remained. Duplicate supplier–tender pairs were also removed.

The observations were organized chronologically into training (2015–2018), validation (2019), and test (2020–2021) periods. This temporal partition replaces the random split used in the original study and prevents future procurement observations from being used during model training.

B. Heterogeneous graph construction

The procurement data were represented as a heterogeneous graph containing Company and Tender nodes. Each observed supplier–tender record forms an edge between the corresponding company and tender. Company and tender nodes were represented in a unified graph index, while node features were constructed from procurement activity and normalized using statistics derived exclusively from the training period. Observed relationships were distinguished from synthetic relationships introduced for anomaly experiments.

C. Temporal split and leakage prevention

Leakage prevention was applied at several levels. Tender observations were separated chronologically, feature normalization used training-period statistics only, and anomaly subjects in validation and test were selected independently from the training anomaly subjects. Synthetic anomaly injection was also performed separately within each temporal partition rather than before the temporal split. Injection audits verified that synthetic relationships did not duplicate observed relationships or violate their designated temporal constraints.

D. Synthetic Anomaly Scenarios

Because the dataset does not contain complete bidder-participation information or verified collusion labels, synthetic structural anomalies were used as controlled experimental ground truth. Two patterns were evaluated: cartel-style coordinated participation and temporal burst participation.

For the cartel scenario, 50 anomaly companies were organized into 10 groups of five companies. Hard, moderate, and easy scenarios introduced one, two, and three synthetic relationships per anomaly company, respectively. The temporal burst scenario modeled concentrated participation by an individual company within 90-, 180-, and 365-day windows, corresponding to the hard, moderate, and easy settings. Anomaly subjects and synthetic relationships were generated independently across temporal partitions. These synthetic labels represent experimental structural patterns and are not evidence of confirmed collusion.

E. GCN and GraphSAGE

GCN and GraphSAGE were implemented using PyTorch Geometric with comparable architectures and optimization settings. Both models use two graph layers, a hidden dimension of 32, ReLU activation, dropout, and two output classes. GCN uses GCNConv, while GraphSAGE uses SAGEConv with mean aggregation. Both models were trained using Adam with class-weighted cross-entropy to address the severe anomaly-class imbalance.

Validation PR-AUC was used for checkpoint selection and early stopping. Both models were evaluated using the same graph representation, temporal partitions, anomaly labels, and training configuration to ensure a controlled comparison.

F. Evaluation protocol

Performance was evaluated using Accuracy, Precision, Recall, F1-Score, ROC-AUC, and PR-AUC. Three random seeds were used across three anomaly difficulty levels, resulting in 18 multi-seed experiments. Additional evaluations included true inductive testing on unseen anomaly subjects, cross-pattern generalization from cartel-style to temporal-burst anomalies, and computational measurements of training and inference time.

For the EWS evaluation, the operating threshold was determined exclusively from the validation population and then frozen for the test evaluation. The test set was not used for threshold optimization.

G. Early Warning System

The selected GraphSAGE model was used to generate company-level anomaly scores. The EWS threshold was calibrated using the validation set and subsequently applied without modification to the unseen test population. Companies were also ranked according to their anomaly scores, and Precision, Recall, and Lift were used to evaluate the screening capability.

The EWS is intended as a prioritization mechanism for further investigation. An anomaly score or alert does not constitute evidence of procurement collusion or procurement misconduct.

4. RESULTS AND DISCUSSION

A. Dataset characteristics

The temporal characteristics of the processed procurement dataset are summarized in Figure 1. The observations were divided chronologically into three temporal partitions, namely training (2015–2018), validation (2019), and testing (2020–2021). The training partition contains 22,623 supplier–tender observations involving 6,865 companies and 16,535 tenders. The validation partition contains 4,347 observations involving 2,522 companies and 3,088 tenders, while the test partition contains 4,372 observations involving 2,632 companies and 2,872 tenders.

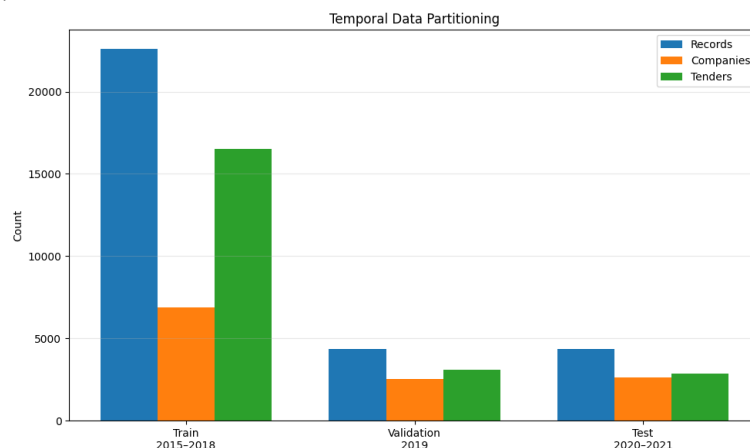


Figure 1. Temporal partitioning of procurement observations across training, validation, and test periods.

The temporal partition preserves the chronological development of the procurement network. The number of companies and tenders therefore differs across periods, reflecting changes in the observed procurement network rather than random sampling. In addition, companies may appear in more than one temporal partition because a supplier can remain active across multiple years. This property is important for evaluating both temporal generalization and the behavior of the models when previously unseen companies emerge in subsequent periods.

Figure 2 presents the class distribution used in the anomaly experiments. Each temporal partition contains 50 synthetic anomaly companies, while the remaining companies are treated as normal for the corresponding experimental scenario. Consequently, the training partition contains 6,815 normal companies and 50 anomaly companies, the validation partition contains 2,472 normal companies and 50 anomaly companies, and the test partition contains 2,582 normal companies and 50 anomaly companies.

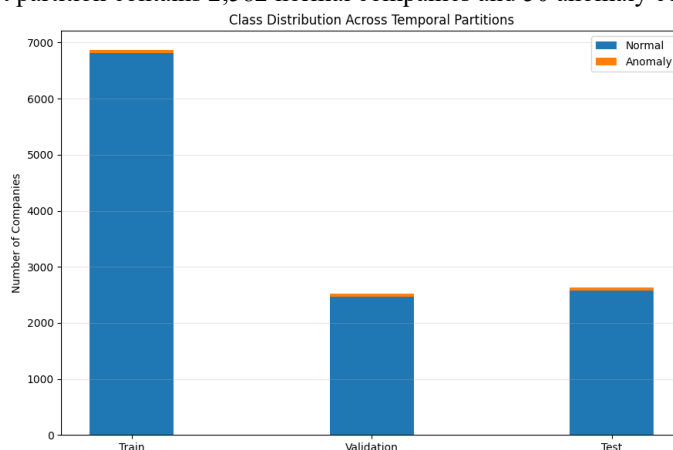


Figure 2. Class distribution across the temporal partitions.

The resulting anomaly prevalence is substantially lower than the normal class, illustrating the severe class imbalance inherent in the experimental setting. The anomaly proportion is approximately 0.73% in the training population, 1.98% in the validation population, and 1.90% in the test population. This imbalance motivates the use of class-weighted loss during model training and the inclusion of PR-AUC, Precision, Recall, and F1-Score in the evaluation protocol rather than relying on accuracy alone.

B. Anomaly scenarios

Because the procurement dataset does not provide a complete record of all firms participating in each tender, it cannot directly establish ground-truth collusion relationships from bidder participation. Therefore, controlled synthetic anomaly injection was used to construct experimental ground truth. The injected relationships represent predefined structural patterns and are not interpreted as evidence of actual procurement collusion.

Two structural anomaly patterns were evaluated: cartel-style co-participation and temporal burst participation. The cartel scenario models groups of companies that exhibit repeated structural relationships with a set of procurement opportunities. For each temporal partition, 10 cartel groups were generated, with five companies in each group. Each group was associated with 10 designated tenders, resulting in 100 synthetic company–tender assignments per scenario before the difficulty-specific sampling was applied.

Three difficulty levels were used to control the strength of the injected structural signal. In the hard scenario, each anomaly company received one synthetic relationship, resulting in 100 injected relationships per temporal partition. The moderate scenario used two synthetic relationships per company, resulting in 200 injected relationships, while the easy scenario used three synthetic relationships per company, resulting in 300 injected relationships.

A second anomaly pattern was introduced to evaluate whether the models could recognize a structurally different anomaly configuration. The temporal burst scenario represents an unusually concentrated participation pattern involving a single company within a specified temporal window. The hard, moderate, and easy scenarios introduced three relationships within 90 days, five relationships within 180 days, and seven relationships within 365 days, respectively. This resulted in 150, 250, and 350 synthetic relationships per temporal partition.

The anomaly subjects for validation and testing were selected independently from the training anomaly subjects. Synthetic relationships were also generated independently within each temporal partition. Audits confirmed that the injected relationships did not duplicate existing supplier–tender relationships and that the temporal burst injections satisfied the specified temporal constraints.

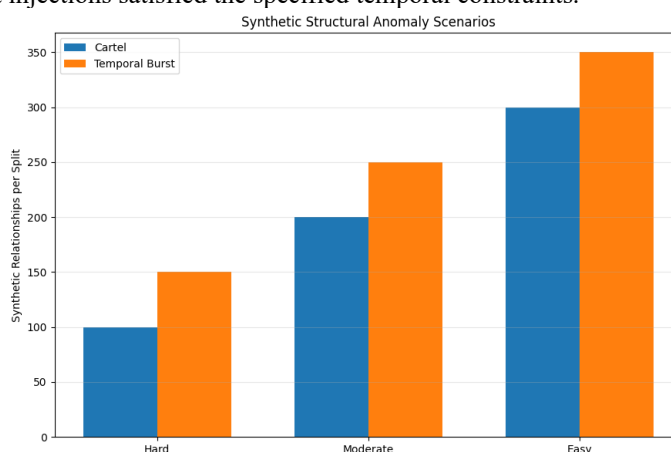


Figure 3. Number of synthetic relationships introduced under different anomaly patterns and difficulty levels.

C. Multi-seed robustness

The robustness of the two models was evaluated across three random seeds (42, 43, and 44) and three anomaly difficulty levels. The resulting 18 experiments provide a broader assessment of model behavior than a single experimental run. Unless otherwise stated, the reported values are presented as mean $\pm$ standard deviation across repeated experimental runs.

As shown in Figure 4, GraphSAGE achieved a mean accuracy of 0.865 ± 0.057 , compared with 0.476 ± 0.279 for GCN. More importantly for the highly imbalanced anomaly-detection task, GraphSAGE obtained a mean F1-Score of 0.176 ± 0.037 and PR-AUC of 0.105 ± 0.007 , while GCN obtained 0.066 ± 0.026 and 0.045 ± 0.015 , respectively. GraphSAGE also achieved a higher mean ROC-AUC of 0.913 ± 0.013 compared with 0.739 ± 0.083 for GCN.

The recall results show a different pattern. GCN obtained a higher mean recall (0.836 ± 0.193) than GraphSAGE (0.742 ± 0.243). However, this higher recall was accompanied by substantially lower precision, indicating that GCN generated more false-positive classifications. This difference is reflected in the lower F1-Score and PR-AUC obtained by GCN.

Across the nine matched model comparisons, GraphSAGE obtained higher F1-Score, ROC-AUC, and PR-AUC in all nine runs. These results provide empirical evidence that the two aggregation mechanisms

behaved differently under the evaluated anomaly scenarios. The results should be interpreted as an empirical comparison under the present experimental configuration rather than as a universal superiority claim for one architecture.

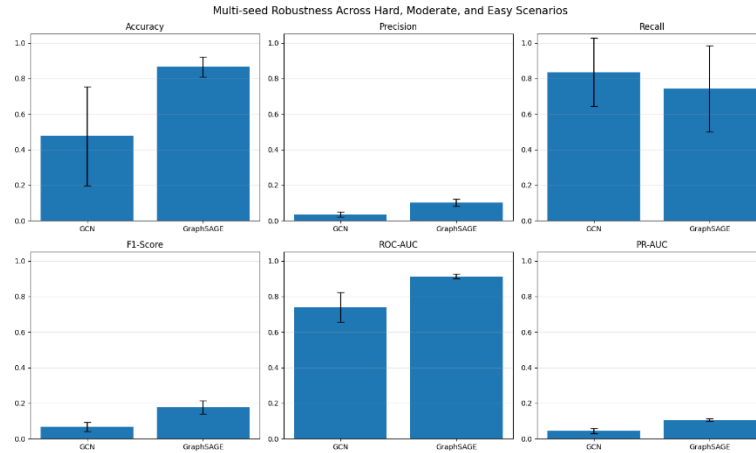


Figure 4. Multi-seed performance of GCN and GraphSAGE across hard, moderate, and easy anomaly scenarios. Error bars represent standard deviation across the three random seeds.

D. Inductive Evaluation on Unseen Anomaly Subjects

To examine whether the models could generalize to anomaly subjects that were not previously used during training, an inductive evaluation was conducted under the moderate cartel scenario. The validation and test anomaly companies were selected independently from the training anomaly subjects. Because companies may persist across multiple temporal periods, this evaluation specifically assesses generalization to previously unseen anomaly subjects rather than requiring all company nodes in the evaluation periods to be completely absent from the training graph.

Figure 5 shows that GraphSAGE achieved an average F1-Score of 0.442 ± 0.125 , compared with 0.151 ± 0.055 for GCN. GraphSAGE also obtained higher mean ROC-AUC and PR-AUC values, with 0.951 ± 0.018 and 0.386 ± 0.007 , respectively, compared with 0.772 ± 0.080 and 0.125 ± 0.033 for GCN.

The recall values were relatively close, with GraphSAGE obtaining 0.880 ± 0.053 and GCN obtaining 0.860 ± 0.242 . The difference in F1-Score was therefore primarily associated with the substantially higher precision obtained by GraphSAGE. This indicates that GraphSAGE produced fewer false-positive classifications while maintaining a similar level of anomaly recall in this evaluation.

The inductive experiment provides an additional evaluation dimension because the anomaly subjects in the evaluation periods were not reused from the training anomaly subjects. Nevertheless, the result remains specific to the experimental procurement graph and synthetic anomaly construction and should not be interpreted as evidence of performance on real-world confirmed collusion cases.

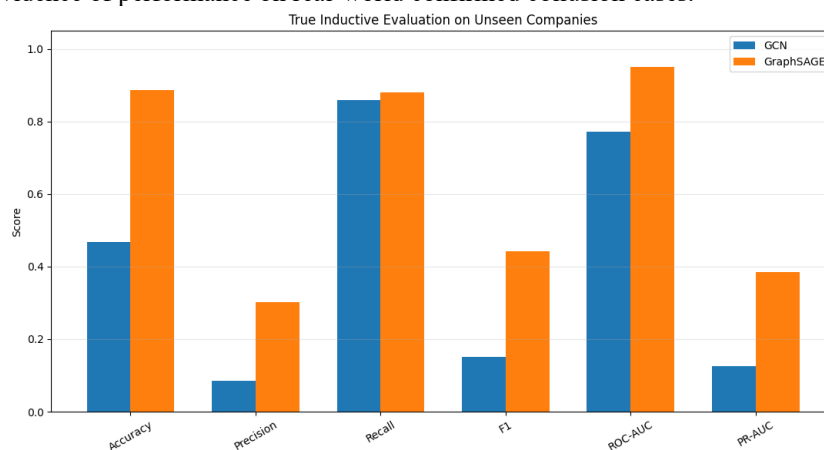


Figure 5. GCN and GraphSAGE performance under true inductive evaluation using unseen anomaly companies.

E. Cross-pattern generalization

A cross-pattern experiment was conducted to evaluate whether the learned representations could generalize beyond the structural anomaly pattern used during training. In this experiment, the models were

trained using the moderate cartel-style anomaly pattern and subsequently evaluated on the temporal burst anomaly pattern.

Figure 6 presents the resulting performance. Under the cross-pattern setting, GCN obtained an F1-Score of 0.249 and PR-AUC of 0.342, while GraphSAGE obtained an F1-Score of 0.358 and PR-AUC of 0.258. GraphSAGE therefore maintained a higher F1-Score in this particular experiment, whereas GCN obtained a higher PR-AUC.

The difference between the metrics illustrates that cross-pattern generalization cannot be adequately characterized using a single performance measure. While GraphSAGE produced a higher threshold-dependent F1-Score in the reported experiment, the ranking-oriented PR-AUC was higher for GCN. This result also demonstrates that a model trained on one synthetic structural pattern may respond differently when evaluated against a structurally different anomaly pattern.

Therefore, the cross-pattern experiment provides a more demanding evaluation than testing and training on identical anomaly configurations. The result should be interpreted as evidence of behavior under the specific cartel-to-temporal-burst transfer setting rather than as a general claim of anomaly-pattern invariance.

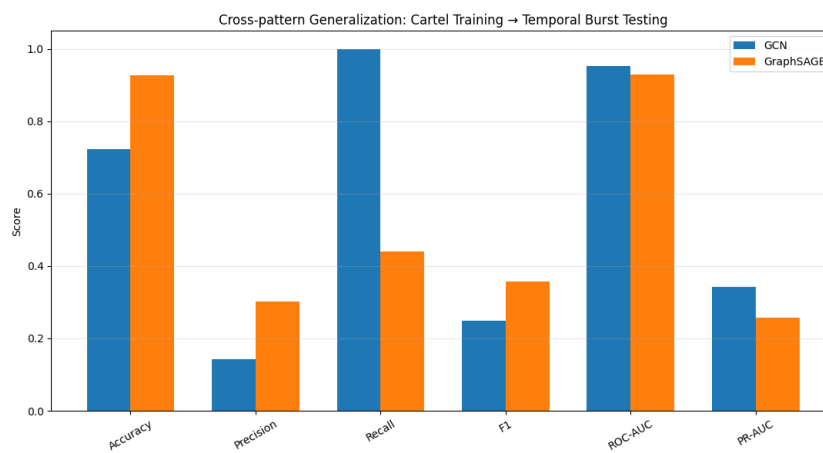


Figure 6. Cross-pattern generalization from cartel-style anomaly training to temporal burst anomaly testing.

F. EWS results

The final experiment evaluated the use of the GraphSAGE model as an Early Warning System (EWS) for structural anomaly screening. The EWS threshold was calibrated exclusively using the validation population. The selected threshold was 0.392749, corresponding to a validation precision of 0.438, recall of 0.920, and F1-Score of 0.594.

The calibrated threshold was subsequently frozen and applied to the unseen test population. As shown in Figure 7, the EWS achieved an accuracy of 0.927, precision of 0.347, recall of 0.660, and F1-Score of 0.455. The threshold-independent metrics were ROC-AUC of 0.931 and PR-AUC of 0.264.

The difference between validation and test performance is expected because the final test evaluation uses a different temporal period and a different anomaly configuration. In particular, the test population contains temporal burst anomalies involving companies that were not used as anomaly subjects during training.

The EWS was also evaluated as a ranking mechanism rather than only as a binary classifier. Figure 8 presents Precision@K and Recall@K for different screening capacities. At Top-50, the system identified 13 of the 50 synthetic anomaly companies, corresponding to a precision of 0.260 and recall of 0.260. At Top-100, 34 anomaly companies were identified, producing a precision of 0.340 and recall of 0.680. At Top-250, recall increased to 0.920, while precision decreased to 0.184.

Figure 9 presents the corresponding screening lift. The highest lift was obtained at Top-100, with a lift of approximately 7.41. This indicates that the proportion of synthetic anomaly companies among the top 100 ranked companies was substantially higher than their prevalence in the complete test population.

These ranking results demonstrate the intended screening-oriented role of the EWS. The system can be used to prioritize a subset of companies for further examination when investigative resources are limited. However, a high anomaly score or a high ranking position does not constitute evidence that a company has engaged in procurement collusion. The output represents a model-derived structural anomaly signal based on the synthetic experimental ground truth.

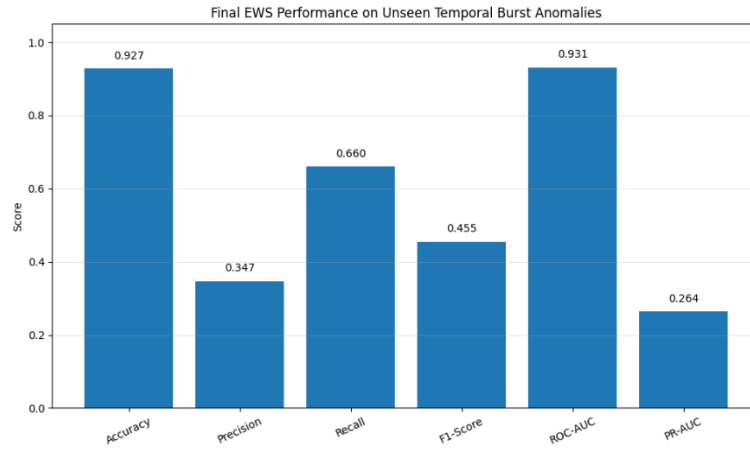


Figure 7. Final EWS performance on unseen temporal burst anomalies using a validation-calibrated threshold.

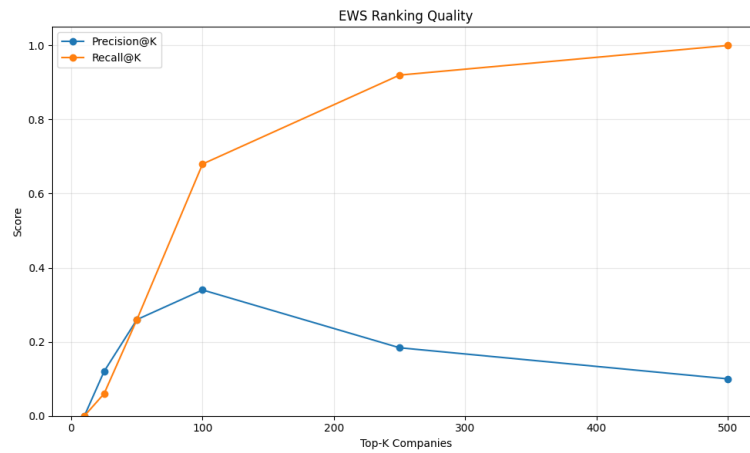


Figure 8. Precision@K and Recall@K of the EWS at different screening capacities.

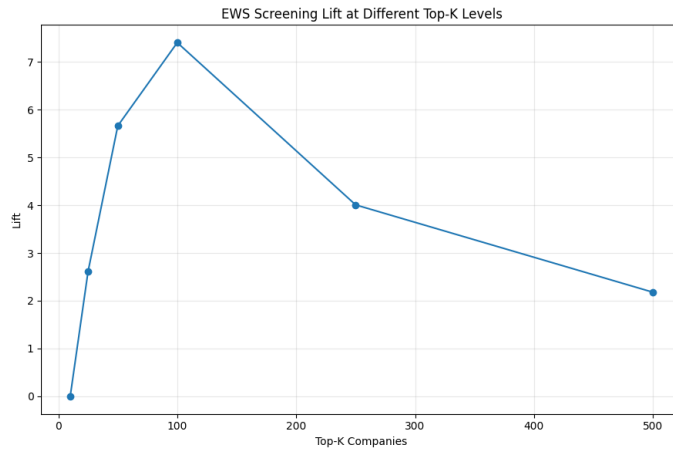


Figure 9. Screening lift achieved by the EWS at different Top-K levels.

G. Discussion

The experimental results provide several important observations regarding the use of graph neural networks for structural anomaly detection in procurement networks. In contrast to the original evaluation, which produced perfect classification scores under a single synthetic anomaly configuration, the revised experiments evaluate the models across multiple random seeds, anomaly difficulty levels, unseen anomaly subjects, and a structurally different anomaly pattern. The results therefore provide a more comprehensive view of model behavior under a highly imbalanced and temporally evolving procurement network.

Across the multi-seed experiments, GraphSAGE obtained higher mean F1-Score, ROC-AUC, and PR-AUC than GCN. GraphSAGE achieved an F1-Score of 0.176 ± 0.037 , ROC-AUC of 0.913 ± 0.013 , and PR-AUC of 0.105 ± 0.007 , whereas GCN achieved 0.066 ± 0.026 , 0.739 ± 0.083 , and 0.045 ± 0.015 , respectively. However, GCN obtained higher mean recall, with 0.836 ± 0.193 compared with 0.742 ± 0.243

(Danny Mario Tua Panjaitan)

for GraphSAGE. The difference indicates that the two models exhibit different precision–recall behavior under the highly imbalanced anomaly setting. The higher recall obtained by GCN was accompanied by substantially lower precision, resulting in a lower F1-Score. Therefore, the results should not be interpreted using recall or accuracy alone.

The class imbalance is an important factor in interpreting these results. The anomaly class represents only a small fraction of the company population, with anomaly prevalence below 2% in the evaluated temporal partitions. Under such conditions, a model can obtain relatively high accuracy while still producing a large number of false-positive or false-negative classifications relative to the minority class. This explains why PR-AUC, precision, recall, and F1-Score are particularly relevant to the present task. The use of class-weighted cross-entropy during training was intended to reduce the dominance of the normal class, while validation PR-AUC was used for model checkpoint selection to provide a more appropriate criterion for the imbalanced anomaly-detection problem.

The true inductive evaluation provides additional evidence regarding the behavior of the models on previously unseen anomaly companies. GraphSAGE obtained an average F1-Score of 0.442 ± 0.125 , compared with 0.151 ± 0.055 for GCN. GraphSAGE also achieved higher ROC-AUC and PR-AUC values, at 0.951 ± 0.018 and 0.386 ± 0.007 , respectively, compared with 0.772 ± 0.080 and 0.125 ± 0.033 for GCN. The recall values were relatively close, with GraphSAGE achieving 0.880 ± 0.053 and GCN achieving 0.860 ± 0.242 . The substantially higher precision obtained by GraphSAGE consequently resulted in a higher F1-Score. These findings suggest that the neighborhood aggregation mechanism used by GraphSAGE can produce useful representations when the anomaly subjects in the evaluation population have not previously been used as anomaly subjects during training. Nevertheless, this result remains specific to the experimental graph and should not be generalized to real-world confirmed collusion without additional external validation.

The cross-pattern experiment provides a further perspective on generalization. The models were trained using the moderate cartel-style anomaly pattern and evaluated on a different temporal burst anomaly pattern. Under this setting, GCN obtained an F1-Score of 0.249 and PR-AUC of 0.342, while GraphSAGE obtained an F1-Score of 0.358 and PR-AUC of 0.258. Thus, GraphSAGE produced the higher F1-Score, whereas GCN produced the higher PR-AUC in this particular cross-pattern experiment. This difference demonstrates that model behavior depends on the evaluation criterion and anomaly structure. It also indicates that performance on one synthetic anomaly pattern should not automatically be interpreted as evidence that the model can detect all forms of structural anomaly. Cross-pattern evaluation is therefore useful for identifying whether a model's learned representation remains informative when the structural characteristics of the anomaly change.

The computational measurements provide another dimension for comparing the architectures. Across the multi-seed experiments, GraphSAGE required an average training time of approximately 4.07 seconds, compared with 4.27 seconds for GCN, while average inference time was approximately 0.020 seconds for GraphSAGE and 0.028 seconds for GCN. In the true inductive evaluation, GraphSAGE required approximately 4.17 seconds for training and 0.022 seconds for inference, compared with 3.21 seconds and 0.025 seconds for GCN. These measurements show that the computational difference between the architectures was relatively small in the present experimental environment. Consequently, the results do not support interpreting GraphSAGE's behavior solely in terms of computational efficiency. The primary differences observed in this study are more directly related to their empirical predictive behavior under the evaluated graph and anomaly configurations.

The final EWS experiment illustrates how the model output can be transformed into an operational screening mechanism. Using the validation population, the operating threshold was calibrated at 0.392749 and subsequently frozen before being applied to the test population. On the unseen test population containing synthetic temporal burst anomalies, the EWS achieved an accuracy of 0.927, precision of 0.347, recall of 0.660, F1-Score of 0.455, ROC-AUC of 0.931, and PR-AUC of 0.264. The reduction in threshold-dependent performance relative to validation is important because it demonstrates that calibration performance does not automatically transfer unchanged to an unseen temporal period and anomaly configuration. This behavior supports the use of temporally separated validation and test populations when evaluating an EWS.

The ranking results further demonstrate the potential use of the model as a prioritization mechanism. At Top-50, the EWS identified 13 of the 50 synthetic anomaly companies, corresponding to a precision of 0.260 and recall of 0.260. At Top-100, it identified 34 anomaly companies, resulting in a precision of 0.340 and recall of 0.680. At Top-250, recall increased to 0.920 while precision decreased to 0.184. The highest screening lift was observed at Top-100, with a lift of approximately 7.41. These results illustrate the trade-off between investigation capacity and anomaly coverage. A smaller screening set may reduce the number of entities requiring manual examination but may also miss a larger proportion of anomalies, whereas expanding the screening set increases coverage at the cost of examining more companies.

From an application perspective, these results suggest that a graph-based EWS can be formulated as a prioritization layer for procurement monitoring. The model can assign an anomaly score to companies and rank them according to their structural characteristics within the procurement network. Such a system could potentially support investigators in allocating limited analytical resources toward companies or relationships exhibiting stronger structural anomaly signals. However, the interpretation of these scores must remain strictly within the scope of anomaly screening. A high score, classification as an anomaly, or high ranking position does not establish that a company participated in bid rigging, collusion, or any other procurement violation.

Several limitations must also be considered when interpreting the findings. First, the GeBIZ dataset used in this study records supplier–tender award observations and does not contain a complete record of all firms that participated in each tender. Consequently, the observed relationships cannot be interpreted as a complete bidding-participation network. Second, the anomaly ground truth used for the experiments is generated synthetically because the dataset does not provide a sufficiently complete set of confirmed collusion labels for direct supervised evaluation. The synthetic patterns therefore provide controlled experimental conditions but do not constitute evidence of actual collusion. Third, the present experiments do not provide an external validation against confirmed real-world collusion cases. Fourth, the experiments are based on a single procurement dataset covering Singapore's GeBIZ procurement observations from 2015 to 2021, and therefore the findings may not directly transfer to procurement systems with different institutional, temporal, or network characteristics. Finally, the cartel-style and temporal-burst scenarios represent predefined structural patterns and cannot encompass the full diversity of potential procurement anomalies.

Overall, the revised experiments indicate that evaluating graph neural networks for procurement anomaly detection requires more than reporting classification accuracy from a single synthetic scenario. The combination of temporal separation, leakage controls, multiple anomaly difficulties, multi-seed evaluation, unseen-company testing, cross-pattern evaluation, and EWS-oriented ranking provides a more demanding assessment of model behavior. Within these experimental conditions, GraphSAGE showed higher F1-Score, ROC-AUC, and PR-AUC across the multi-seed comparisons, while GCN demonstrated higher average recall. The EWS experiment further shows how model scores can be used for ranked structural-anomaly screening. These findings support the feasibility of graph-based anomaly screening while also emphasizing the need for external validation and real-world ground truth before such systems can be used to make substantive determinations regarding procurement misconduct.

5. CONCLUSION

This study has several limitations that should be considered when interpreting the experimental findings. First, the Government Procurement via GeBIZ dataset records observed supplier–tender award relationships but does not provide a complete record of all companies participating in each tender. Therefore, the constructed procurement network should not be interpreted as a complete bidding-participation network. Second, because the dataset does not provide a sufficiently complete set of confirmed collusion labels, the anomaly ground truth in this study was generated through controlled synthetic structural anomaly injection. The synthetic anomalies enable reproducible evaluation under different difficulty levels and structural configurations, but they do not constitute evidence of actual procurement collusion. Third, the experiments do not include external validation against confirmed real-world collusion cases. Consequently, the reported results should be interpreted as evidence of model behavior under controlled experimental conditions rather than as evidence that the models can directly identify legally confirmed procurement violations. Fourth, the study uses procurement observations from a single GeBIZ dataset covering the 2015–2021 period, and the findings may therefore not directly generalize to procurement networks with different institutional, temporal, or structural characteristics. Finally, the cartel-style and temporal-burst scenarios represent predefined anomaly configurations and cannot capture the full diversity of structural patterns that may occur in real procurement environments.

Despite these limitations, the revised experimental framework provides a more rigorous evaluation of GCN and GraphSAGE for structural anomaly detection in procurement networks. The use of temporal partitioning, leakage prevention, multiple anomaly difficulty levels, multi-seed evaluation, true inductive testing, and cross-pattern generalization addresses several methodological limitations of a single random split and single synthetic anomaly configuration. Across the multi-seed experiments, GraphSAGE achieved higher mean F1-Score, ROC-AUC, and PR-AUC than GCN, while GCN achieved higher mean recall. The true inductive evaluation further showed that GraphSAGE maintained higher F1-Score, ROC-AUC, and PR-AUC when evaluated on previously unseen anomaly companies under the moderate cartel scenario. The cross-pattern experiment demonstrated that model behavior can change when the structural anomaly pattern differs from the pattern used during training, emphasizing the importance of evaluating anomaly detectors beyond same-pattern testing.

The final EWS experiment demonstrated how the GraphSAGE output can be transformed into a screening-oriented mechanism. Using a threshold calibrated exclusively on the validation population, the EWS achieved an F1-Score of 0.455, ROC-AUC of 0.931, and PR-AUC of 0.264 on the unseen test population containing synthetic temporal-burst anomalies. The ranking analysis further showed that the Top-100 screening level achieved a lift of approximately 7.41, while identifying 68% of the synthetic anomaly companies. These results illustrate the potential of graph-based models to support prioritization when investigative resources are limited. However, the EWS should be regarded as a structural anomaly screening mechanism rather than a system for establishing or confirming collusion.

In conclusion, the findings indicate that graph neural networks can provide useful representations for detecting experimentally constructed structural anomalies in temporal procurement networks. The comparative results show distinct precision-recall behavior between GCN and GraphSAGE, with GraphSAGE demonstrating stronger F1-Score and ranking-oriented discrimination across the multi-seed experiments, while GCN maintained higher average recall. More importantly, the revised evaluation demonstrates that procurement anomaly detection should be assessed under realistic challenges including severe class imbalance, temporal separation, unseen entities, and changes in anomaly structure. Future research should therefore focus on obtaining or constructing validated real-world ground-truth cases, incorporating more complete bidder-participation information, evaluating additional procurement datasets, and testing a broader range of structural anomaly patterns. Such extensions are necessary before the proposed approach can be evaluated for deployment in operational procurement monitoring environments.

REFERENSI

- Baránek, B., Musolff, L., & Titl, V. (2021). *Detection of collusive networks in e-procurement*. Utrecht University. https://www.uu.nl/sites/default/files/LEG_USE_WP_21-11.pdf
- Bejger, S. (2024). Machine learning in cartel screening—The case of parallel pricing on the Polish retail gasoline market. *Energies*, 17(16), 4184. <https://doi.org/10.3390/en17164184>
- Hamilton, W. L., Ying, R., & Leskovec, J. (2017). Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems* (Vol. 30). <https://doi.org/10.48550/arXiv.1706.02216>
- Kim, H., Lee, B. S., Shin, W.-Y., & Lim, S. (2022). *Graph anomaly detection with graph neural networks: Current status and challenges*. arXiv. <https://doi.org/10.48550/arXiv.2209.14930>
- Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. *International Conference on Learning Representations*. <https://doi.org/10.48550/arXiv.1609.02907>
- Li, M., & Walsh, J. (2024). FEDGAT-DCNN: Advanced credit card fraud detection using federated learning, graph attention networks, and dilated convolutions. *Electronics*, 13(16), 3169. <https://doi.org/10.3390/electronics13163169>
- Liu, Y., Pan, S., Wang, Y. G., Xiong, F., Wang, L., Chen, Q., & Lee, V. C. (2023). Anomaly detection in dynamic graphs via transformer. *IEEE Transactions on Knowledge and Data Engineering*, 35(12), 12081–12094. <https://doi.org/10.1109/TKDE.2021.3124061>
- Ma, X., Wu, J., Xue, S., Yang, J., Zhou, C., Sheng, Q. Z., Xiong, H., & Akoglu, L. (2023). A comprehensive survey on graph anomaly detection with deep learning. *IEEE Transactions on Knowledge and Data Engineering*, 35(12), 12012–12038. <https://doi.org/10.1109/TKDE.2021.3118815>
- Motie, S., & Raahemi, B. (2024). Financial fraud detection using graph neural networks: A systematic review. *Expert Systems with Applications*, 240, 122156. <https://doi.org/10.1016/j.eswa.2023.122156>
- Pourhabibi, T., Ong, K.-L., Kam, B. H., & Boo, Y. L. (2020). Fraud detection: A systematic literature review of graph-based anomaly detection approaches. *Decision Support Systems*, 133, 113303. <https://doi.org/10.1016/j.dss.2020.113303>
- Rashidi, A., Tamošaitienė, J., Ravanshadnia, M., & Sarvari, H. (2023). A scientometric analysis of construction bidding research activities. *Buildings*, 13(1), 220. <https://doi.org/10.3390/buildings13010220>
- Schlichtkrull, M., Kipf, T. N., Bloem, P., van den Berg, R., Titov, I., & Welling, M. (2018). Modeling relational data with graph convolutional networks. In A. Gangemi, R. Navigli, M.-E. Vidal, P. Hitzler, R. Troncy, L. Hollink, A. Tordai, & M. Alam (Eds.), *The Semantic Web* (pp. 593–607). Springer. https://doi.org/10.1007/978-3-319-93417-4_38
- Wu, B., Chao, K.-M., & Li, Y. (2024). Heterogeneous graph neural networks for fraud detection and explanation in supply chain finance. *Information Systems*, 121, 102335. <https://doi.org/10.1016/j.is.2023.102335>
- Xu, F., Wang, N., Wu, H., Wen, X., Zhao, X., & Wan, H. (2024). Revisiting graph-based fraud detection in sight of heterophily and spectrum. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(8), 9214–9222. <https://doi.org/10.1609/aaai.v38i8.28773>
- Zeng, Y. & Tang, J. (2021). RLC-GNN: An improved deep architecture for spatial-based graph neural network with application to fraud detection. *Applied Sciences*, 11(12), 5656. <https://doi.org/10.3390/app11125656>