APPLICATIONS OF GRAPH NEURAL NETWORKS IN OUTLIER DETECTION

Thesis Submitted
in Partial Fulfillment of the Requirements for the
Degree of

MASTER OF TECHNOLOGY

in

ARTIFICIAL INTELLIGENCE

by

ROHIT SAINI

(2K22/AFI/18)

Under the Supervision of

Dr. ANURAG GOEL
Assistant Professor, Department of Computer Science and Engineering
Delhi Technological University

Department of Computer Science and Engineering

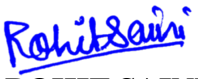
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi 110042

June, 2024

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

ACKNOWLEDGMENT

I wish to express my sincerest gratitude to **Dr. Anurag Goel** for his continuous guidance and mentorship that he provided during research work. He showed me the path to achieving targets by explaining all the tasks to be done and explained to me the importance of this work as well as its industrial relevance. He was always ready to help me and clear our doubts regarding any hurdles in this project. Without his constant support and motivation, this work would not have been successful.



Place: Delhi **ROHIT SAINI**

Date:

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

CANDIDATE'S DECLARATION

I, **Rohit Saini**, **2K22/AFI/18**, of **M.Tech. (AI)**, hereby certify that the work which is being presented in the thesis entitled "**Applications of Graph Neural Networks in Outlier Detection**" in partial fulfillment of the requirement for the award of the degree of Master of Technology, submitted in the Department of Computer Science and Engineering, Delhi Technological University is an authentic record of my own work carried out during the period from to under the supervision of **Dr. Anurag Goel**. The matter presented in the thesis has not been submitted by me for the award of any other degree of this or any other institute.

Candidate's Signature

This is to certify that the student has incorporated all the corrections suggested by the examiners in the thesis and the statement made by the candidate is correct to the best of our knowledge.

Signature of Supervisor **Signature of External Examiner**

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

CERTIFICATE BY THE SUPERVISOR

Certified that **Rohit Saini (2K22/AFI/18)** has carried out their research work presented in this thesis entitled "**Applications of Graph Neural Networks in Outlier Detection**" for the award of **Master of Technology** from Computer Science and Engineering, Delhi Technological University, Delhi under my supervision. The thesis embodies results of original work, and studies are carried out by the student himself and the contents of the thesis do not form the basis for the award of any other degree to the candidate or to anybody else from this or any other University/Institution.

(Dr. ANURAG GOEL)

(Assistant Professor)
(Department of Computer Science and Engineering)
(Delhi Technological University)

Date:

Abstract

Graph Neural Networks (GNNs) have become a tool, in detecting outliers within graphs. When designing GNNs a key aspect is choosing a filter that suits the task. This research delves into outlier analysis by examining the graph spectrum and presents a finding; the presence of outlier leads to a 'right shift' effect, where the energy distribution in the spectrum moves towards frequencies. This revelation carries implications for GNN design suggesting that conventional low pass filters may not be ideal, for outlier detection. To address this challenge, we propose the Beta Wavelet Graph Neural Network (BWGNN), which incorporates spectral and spatial localized band-pass filters. These filters are specifically designed to handle the 'right-shift' phenomenon, providing a more effective approach to outlier detection. We evaluate the performance of BWGNN on four large scale outlier detection datasets and demonstrate its superiority over existing methods. Our findings not only shed light on the spectral properties of graph outliers but also pave the way for more sophisticated GNN architectures that can better capture the nuances of anomalous behavior in graph data.

Contents

List of Tables

List of Figures

List of Symbols

$h_v^{(k+1)}$	Node feature vector at layer $k+1$
$W^{(k)}$	Weight matrix at layer k
AGG	Aggregation function
$h_u^{(k)}$	Neighbor node feature vector at layer k
$\mathcal{N}(v)$	Set of neighbors of node v
σ	Activation function
$H^{(l)}$	Input feature matrix at layer l
$H^{(l+1)}$	Output feature matrix at layer $l+1$
$\tilde{A}$	Adjacency matrix with added self-loops
$\tilde{D}$	Diagonal degree matrix of $\tilde{A}$
α_{ij}	Attention coefficient between node i and node j
a	Learnable attention vector
$\|\cdot\|_2$	L2 norm
U	Matrix of eigenvectors of the Laplacian
Λ	Diagonal matrix of eigenvalues
$g(\Lambda)$	Spectral filter function
$\hat{X}$	Transformed feature matrix
$\hat{Y}$	Filtered feature matrix
Y	Output feature matrix after spectral convolution
$\mathscr{L}_{\textbf{class}}$	Class loss
C_i	Class-specific features of node i
p	Prototype vector
$\mathscr{L}_{\textbf{connectivity}}$	Connectivity loss
S_i	Surrounding features of node i
$\mathscr{L}_{\textbf{GDN}}$	Overall loss in GDN
$\mathscr{L}_{\textbf{classification}}$	Classification loss
λ_1, λ_2	Regularization parameters
$\mathscr{L}_{\textbf{contrastive}}$	Contrastive loss

Chapter 1

INTRODUCTION

1.1 Background

Graph Neural Networks (GNNs) have become essential in the realm of machine learning, for data structured in graphs. These networks are crafted to make the most of the data within graphs and have been effectively used for tasks such, as categorizing nodes predicting links and classifying graphs. The capability of GNNs to grasp graph topology and characteristics makes them well suited for outlier detection, which aims to pinpoint patterns within data that do not conform to expected behavior.

Outlier detection in graphs is a critical task with applications across numerous domains such as cybersecurity, finance, and social network analysis. Traditional outlier detection techniques often rely on statistical methods or shallow machine learning models that may not fully capture the complex dependencies within graph data. GNNs address this limitation by employing deep learning architectures that can learn representations of nodes and edges, considering both their features and the structure of the graph.Despite their success, GNNs face challenges when dealing with outlier in graphs.

Outliers are often sparse and structurally different from normal instances, which can make them difficult to detect using standard GNN architectures. These architectures typically use low-pass filters that smooth features over the graph, which may inadvertently obscure the distinctive signals of outliers. Recent research has highlighted the importance of considering the spectral properties of graphs when designing GNNs for outlier detection. The graph spectrum, derived from the eigenvalues and eigenvectors of the graph Laplacian, provides valuable insights into the structure of the graph.

It has been observed that outliers can cause a 'right-shift' in the graph spectrum, where the energy distribution moves towards higher frequencies. This phenomenon suggests that the design of spectral filters in GNNs should be rethought to effectively capture the high-frequency components associated with outliers. In light of these findings, there is a growing interest in developing new GNN architectures that can better handle the spectral characteristics of graph outliers.

By rethinking the design of spectral filters and incorporating insights from the graph spectrum, researchers aim to improve the performance of GNNs in outlier detection tasks.

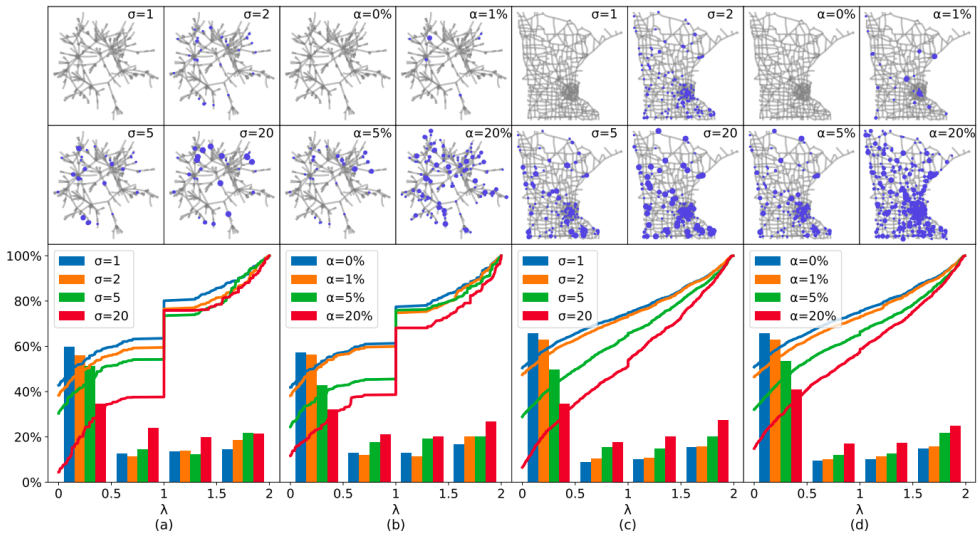


Figure 1.1: The impact of graph outliers is depicted in the spatial domain (top) and spectral domain (bottom) for various anomaly levels[1]

1.2 Problem Statement

The identification of irregularities, in graph based information is an issue in areas such as cybersecurity, fraud prevention and social media assessment. Uncommon occurrences within graphs typically appear as arrangements or deviations from the pattern potentially signaling risks or revealing important findings. Conventional techniques for spotting outliers face hurdles due, to the nature and vast size of graph data hence why Graph Neural Networks (GNNs) have been embraced for their capacity to adapt and comprehend complexities. and generalize from graph topology and node features.

However, GNNs are predominantly designed with low-pass spectral filters that emphasize smoothness and gradual feature propagation across the graph. This design paradigm is well-suited for tasks that rely on homophily, where similar nodes are expected to exhibit similar features. Outlier detection, on the contrary, requires the identification of dissimilar and rare patterns that low-pass filters may inadvertently dilute or overlook.

The central problem addressed in this research is the inadequacy of conventional GNN architectures in effectively detecting outliers within graphs. Specifically, the standard low-pass filtering approach fails to account for the 'right-shift' phenomenon observed in the graph spectrum when outliers are present. This shift indicates that outliers are associated with higher-frequency components, which are not adequately captured by existing GNN models.

The challenge lies in rethinking the design of GNNs to incorporate spectral filters that can detect and preserve the high-frequency signals indicative of outliers. The goal is to develop a GNN architecture that not only learns from the graph's structure and features but also adapts to the unique spectral characteristics of outliers, thereby enhancing the detection capabilities and robustness of the model.

1.3 Importance of Graph-Structured Data

In times there has been an increase, in the use of graph structured data especially in areas such as social media, online shopping and financial activities. In these scenarios nodes represent entities like users, products or transactions while edges symbolize connections between these entities such as friendships, joint purchases or monetary transactions. The interconnected nature of graph data offers insights that can enhance the accuracy of detecting outliers. By utilizing graph based strategies we can capture the relationships and structures, within these datasets resulting in sophisticated and efficient outlier detection methods compared to traditional isolated data analysis techniques.

1.4 Challenges in Graph Outlier Detection

Graph Neural Networks (GNNs) have demonstrated tremendous power in processing graph-structured data for rich relational and interaction analysis. GNNs leverage information flow over the graph by aggregating features from neighbor nodes to learn powerful representations. Traditional GNN architectures have shown to be successful in a number of graph-based tasks, nevertheless they suffer from key issues when it comes to outlier detection including vulnerability to over-smoothing and homophily (the tendency for nodes with similar feature vectors to connect) based nature. GNN models collect features from neighbors and over-smoothing happens when aggregated feature do not exhibit enough difference to properly understand the attributes of outliers.

1.5 Structural Distribution Shift

Detecting outliers in graphs poses a challenge due, to something called distribution shift (SDS). This essentially refers to how the distribution of aspects, like node connections, changes between training and testing phases. This shift can really impact how graph network (GNN) models can adapt, especially because outliers tend to have more connections to different types of nodes compared to regular ones. While models, like the Graph Decomposition Network (GDN) try to tackle this issue by adjusting features they often struggle to handle SDS.

1.6 Proposed Solution

In our research we present the Beta Wavelet Graph Neural Network (BWGNN) a method aimed at overcoming the shortcomings of existing outlier detection techniques based on graph networks. BWGNN uses graph theory to tackle the SDS issue by examining the domain features of graph data. Unlike spatial domain methods our approach utilizes filters to effectively address outliers that lead to shifts, in spectral energy distribution. We showcase the effectiveness of BWGNN by conducting experiments, on known datasets like Amazon and YelpChi using a 40% split between test and train data. The results demonstrate that BWGNN outperforms methods in detecting outliers proving its reliability in environments with structural distribution shifts. Our key contributions include:-

Introducing BWGNN, a spectral based Graph Neural Network tailored for outlier detection in graph data. Offering an analysis of the shift in distribution and its influence on outlier detection models based on GNNs. Performing experiments on real world datasets to confirm the effectiveness of BWGNN comparing its performance, with the techniques. Through this study our goal is to advance graph outlier detection by presenting a method that enhances detection accuracy and resilience to structural distribution shifts.

Chapter 2

PRELIMINARIES

2.1 Graph Outlier Detection (GOD)

Many traditional methods, for spotting occurrences tend to miss out on the interconnected nature of data in scenarios, like social networks or transaction records. GOD takes advantage of these connections by treating data as graphs with nodes representing entities and edges showing how they interact. This approach makes it easier to spot outliers that might not stand out when looking at data points but become noticeable when considering their relationships. By taking into account these connections GOD can uncover patterns and irregularities that would escape detection using techniques thereby improving the reliability and precision of outlier detection.

2.2 Structural Distribution Shift (SDS)

Changes, in the structure of the graph during training and testing are known as SDS[2]. These variations in node connections can make it difficult for GNN models to perform well on data. It's important to tackle SDS[3] to ensure outlier detection when the data distribution changes. The changing nature of graph data, with connections and entities appearing regularly amplifies the challenge of dealing with SDS requiring adaptable methods to handle these shifts smoothly.

2.3 Spectral Graph Theory

The field of graph theory delves into understanding a graphs attributes through studying the eigenvalues and eigenvectors of matrices linked to the graph like the matrix. This method offers a glimpse into the features of the graph in terms of frequencies allowing for the creation of filters that can pinpoint particular frequency components tied to irregularities. Through examining the spectral traits of graphs scientists can pinpoint trends and properties that signal outliers thereby boosting GNNs detection abilities[4].

Chapter 3

LITERATURE REVIEW

3.1 Graph Neural Networks (GNNs)

Graph Neural Networks have emerged as a powerful tool for various graph-based tasks, including outlier detection[5]. GNNs leverage the graph structure to aggregate information from a node's neighborhood, which helps in learning robust node representations. However, vanilla GNNs often struggle with outlier detection due to the over-smoothing problem, where the features of neighboring nodes become indistinguishable. This issue is particularly problematic for outliers, as their distinguishing features get averaged out during the aggregation process.

GNN Update Equation:

$$h_v^{(k+1)} = \sigma\left(W^{(k)} \cdot \text{AGG}\left(\{h_u^{(k)} : u \in \mathcal{N}(v)\}\right)\right) \tag{3.1}$$

GCN Layer Update:

$$H^{(l+1)} = \sigma\left(\tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}H^{(l)}W^{(l)}\right) \tag{3.2}$$

3.2 Attention Mechanisms

To mitigate the over-smoothing issue, several approaches have incorporated attention mechanisms into GNNs. These mechanisms allow the model to weigh the importance of different neighbors differently, thus preserving the unique characteristics of anomalous nodes. For instance, methods like GAT (Graph Attention Networks) utilize self-attention layers to focus on the most relevant parts of the neighborhood, improving the detection of outliers[6].

GAT Attention Coefficient:

$$\alpha_{ij} = \frac{\exp\left(\text{LeakyReLU}\left(a^T[Wh_i||Wh_j]\right)\right)}{\sum_{k\in\mathcal{N}(i)}\exp\left(\text{LeakyReLU}\left(a^T[Wh_i||Wh_k]\right)\right)} \tag{3.3}$$

GAT Node Update:

$$h_i' = \sigma\left(\sum_{j\in\mathcal{N}(i)} \alpha_{ij} W h_j\right) \tag{3.4}$$

3.3 Resampling Strategies

Another approach to address the limitations of GNNs in outlier detection involves selective resampling of neighborhood information. Techniques such as Care-GNN adaptively sample neighbors based on their similarity, ensuring that the model aggregates information from more relevant neighbors and reduces the impact of noisy connections[7].

Resampling Process:

$$\mathcal{N}'(i) = \text{Sample}(\mathcal{N}(i), p) \tag{3.5}$$

3.4 Auxiliary Losses

Adding auxiliary losses during training is another strategy to enhance the model's robustness. These losses can be designed to enforce certain properties in the learned representations, such as maintaining high separability between normal and anomalous nodes. For example, some models introduce contrastive losses that encourage the separation of outlier features from those of normal nodes.

Contrastive Loss:

$$\mathcal{L}_{\text{contrastive}} = \sum_{(i,j)\in\mathcal{P}} \|h_i - h_j\|_2^2 - \sum_{(i,k)\in\mathcal{N}} \|h_i - h_k\|_2^2 \tag{3.6}$$

3.5 Spectral Approaches

While most GNN-based outlier detection methods operate in the spatial domain, recent studies have explored the spectral domain for better handling of outliers[8]. The spectral domain analysis focuses on the graph's frequency components, where outliers often induce a 'right-shift' in the spectral energy distribution, concentrating more energy in high frequencies. Techniques like the Beta Wavelet Graph Neural Network (BWGNN) utilize band-pass filters to target these spectral characteristics, effectively distinguishing outliers from normal nodes.

Spectral Convolution:

$$\hat{X} = U^T X \tag{3.7}$$

$$\hat{Y} = g(\Lambda)\hat{X} \tag{3.8}$$

$$Y = U\hat{Y} \tag{3.9}$$

3.6 Addressing Structural Distribution Shift (SDS)

The Graph Decomposition Network (GDN)[9] introduces a novel approach to mitigate SDS by differentiating structural patterns for outliers and normal nodes. GDN separates node features into class-specific and surrounding features. For outliers, it ensures that their critical features remain invariant to the heterophily shift by constraining them through a prototype vector that updates dynamically during training. This approach reduces the influence of heterophilous neighbors and enhances the model's robustness to SDS. For normal nodes, GDN preserves the connectivity features to leverage homophily, thus benefiting from stable neighborhood patterns.

GDN splits node features $\mathbf{X}$ into class features $\mathbf{C}$ and surrounding features $\mathbf{S}$:

$$X = C + S \tag{3.10}$$

Graph Decomposition Network (GDN) Class Loss:

$$\mathscr{L}_{\text{class}} = \sum_i \|C_i - p\|_2^2 \tag{3.11}$$

GDN Connectivity Loss:

$$\mathscr{L}_{\text{connectivity}} = \sum_{(i,j)\in E} \|S_i - S_j\|_2^2 \tag{3.12}$$

GDN Overall Loss:

$$\mathscr{L}_{\text{GDN}} = \mathscr{L}_{\text{classification}} + \lambda_1 \mathscr{L}_{\text{class}} + \lambda_2 \mathscr{L}_{\text{connectivity}} \tag{3.13}$$

Comparative Performance: Experimental evaluations on benchmark datasets like Amazon and YelpChi demonstrate the superiority of GDN over traditional GNN-based models. GDN consistently achieves higher performance metrics in environments with significant SDS, showcasing its robustness and adaptability[10]. This performance is attributed to its ability to maintain the distinct features of outliers while effectively leveraging the stable patterns of normal nodes.

3.7 Graph Outlier Detection (GOD) Methodologies

3.7.1 Semi-Supervised Learning

Graph outlier detection often employs semi-supervised learning methods due to the scarcity of labeled outliers. These methods train on a small set of labeled nodes and use the learned patterns to classify unlabeled nodes. The challenge lies in the imbalance between normal and anomalous nodes, which can bias the model towards the majority class. Techniques like GCN (Graph Convolutional Network) and Graph-SAGE (Graph Sample and Aggregate) are commonly used, but they need enhancements to handle the imbalance and SDS effectively[11].

3.7.2 Autoencoder-Based Methods

Autoencoders have been widely used for outlier detection due to their ability to learn compact representations of normal data[12]. In the context of graphs, methods like DONE (Deep Anomaly Detection on Attributed Networks) use graph autoencoders to learn node representations. These methods minimize the reconstruction error of normal nodes while highlighting outliers. However, their effectiveness is limited when the graph structure significantly changes between training and testing, a scenario often caused by SDS.

Chapter 4

METHODOLOGY

4.1 Structural Distribution Shift in GAD

In dealing with SDS in GAD it's crucial to grasp how changes, in node connectivity and distribution impact the performance of GNN models. Traditional GNNs gather data from neighboring nodes, which may blur the features of outliers especially when they are linked to normal nodes. This blending can result in smoothing causing node representations to blend together and hinder outlier detection. By tackling SDS we can enhance the models capacity to adapt from training data to scenarios ultimately boosting its reliability and precision.

4.2 Beta Wavelet Graph Neural Network (BWGNN)

4.2.1 Design of BWGNN

The design of the Beta Wavelet Graph Neural Network (BWGNN) is a significant advancement in the field of graph neural networks, particularly for the task of outlier detection[13]. BWGNN is designed to address the limitations of traditional GNNs by incorporating spectral and spatial localized band-pass filters that are better suited for handling the 'right-shift' phenomenon observed in outliers[1]. Specifically, BWGNN adopts the following propagation process with Weighted cross-entropy loss is used for the training of BWGNN:

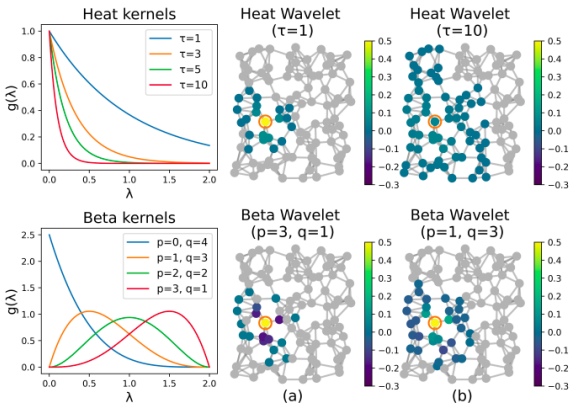


Figure 4.1: Comparative analysis of Heat wavelets and Beta wavelets in both the spectral domain (left) and spatial domain (right)[1].

$$Z_i = W_i \cdot C_{-i}(\text{MLP}(X)) \tag{4.1}$$

$$H = \text{AGG}([Z_0, Z_1, \ldots, Z_C]) \tag{4.2}$$

$$L = \sum_i (\gamma y_i \log(p_i) + (1 - y_i) \log(1 - p_i)) \tag{4.3}$$

where γ is the ratio of outlier labels ($y_i = 1$) to normal labels ($y_i = 0$)[1].

Key Features of BWGNN:

- **Spectral and Spatial Localized Band-Pass Filters:** BWGNN utilizes band-pass filters that are both spectral and spatially localized. This allows the network to focus on specific frequency ranges within the graph spectrum, enhancing its ability to detect outliers that are characterized by higher-frequency components[14].
- **Beta Kernel:** The core of BWGNN's method is the Beta kernel, which addresses higher frequency outliers through flexible, spatial/spectral-localized, and band-pass filters1. This contrasts with the widely used Heat kernels and allows BWGNN to be more effective in identifying outliers.
- **Handling Over-Smoothing:** Traditional GNNs often suffer from the over-smoothing issue when aggregating information from node neighborhoods, which can make outliers less distinguishable. BWGNN's design mitigates this issue by preserving the distinctiveness of anomalous nodes through its specialized filters.

Advantages of BWGNN:

- **Improved outlier Detection:** By focusing on the right-shift phenomenon, BWGNN can detect outliers more effectively than traditional GNNs that rely on low-pass filters.
- **Flexibility:** The Beta kernel provides flexibility in the design, allowing BWGNN to adapt to different types of outliers and graph structures.
- **Efficiency:** BWGNN's localized filters enable it to process graph data more efficiently, making it suitable for large-scale applications.

BWGNN represents a thoughtful reimagining of GNN architecture for outlier detection. Its design leverages the spectral properties of graph outliers to provide a more accurate and efficient tool for identifying irregular patterns within graph-structured data.

4.2.2 Theoretical Justification

The theoretical justification for the design of the Beta Wavelet Graph Neural Network (BWGNN) is rooted in the spectral graph theory and the need to address the unique challenges posed by outlier detection in graph-structured data. Here's an in-depth look at the theoretical underpinnings of BWGNN :-

- **Spectral Graph Theory:** Spectral graph theory provides a framework for analyzing the properties of graphs using the eigenvalues and eigenvectors of matrices associated with the graph, such as the graph Laplacian. The spectrum of a graph reveals important structural information, including the presence of communities, bottlenecks, and outliers.

- **The 'Right-Shift' Phenomenon:** The observation of the 'right-shift' phenomenon, where the spectral energy distribution shifts towards higher frequencies in the presence of outliers, is a key theoretical insight. This shift indicates that outliers are associated with high-frequency components in the graph spectrum, which are not adequately captured by traditional low-pass filters used in GNNs.
- **Localized Band-Pass Filters:** BWGNN employs localized band-pass filters that are capable of capturing these high frequency components. The theoretical justification for this approach is that it allows the network to focus on the spectral regions most affected by outliers, enhancing the model's sensitivity to irregular patterns.
- **Beta Kernel:** The Beta kernel used in BWGNN is theoretically justified by its ability to provide a flexible response to different frequency ranges. Unlike the rigid structure of traditional kernels, the Beta kernel can be adjusted to target specific spectral bands, making it more effective for outlier detection.
- **Balancing Aggregation and Preservation:** A fundamental challenge in GNN design is balancing the aggregation of information from node neighborhoods with the preservation of distinctive node features. BWGNN's theoretical design addresses this by ensuring that the aggregation process does not lead to over-smoothing, which can obscure outliers.
- **Efficiency and Scalability:** The theoretical design of BWGNN also considers the computational efficiency and scalability of the model. By using localized filters, BWGNN reduces the computational complexity, making it suitable for large-scale graphs where outliers need to be detected in real-time.

The theoretical justification for BWGNN lies in its ability to address the limitations of existing GNNs in the context of outlier detection. By leveraging spectral graph theory and the 'right-shift' phenomenon, BWGNN provides a theoretically sound and practically effective solution for detecting outliers in graph-structured data.

4.2.3 Addressing the 'Right-Shift'

Addressing the 'right-shift' in the spectral energy distribution is a crucial aspect of outlier detection in graphs.The 'right-shift' refers to the phenomenon where the presence of outliers in a graph leads to a redistribution of spectral energy towards higher frequencies1. This shift challenges the traditional design of Graph Neural Networks (GNNs), which typically employ low-pass filters that focus on smooth, low-frequency signals and may fail to capture the high-frequency components indicative of outliers.

Strategies for Addressing the 'Right-Shift':

- **Band-Pass Filters:** One approach to address the 'right-shift' is the use of band-pass filters in the design of GNNs. These filters are capable of isolating the frequency bands where outliers are likely to manifest, allowing the network to focus on the relevant spectral components.
- **Spectral Localization:** Spectral localization involves designing filters that are sensitive to specific parts of the graph spectrum. This allows for targeted analysis of the high-frequency regions affected by the 'right-shift', enhancing the detection of outliers[15].

- **Wavelet Theory:** The application of wavelet theory to GNNs provides a framework for creating band-pass filters that are both spectral-localized and spatial localized.This dual localization is key to capturing the nuanced spectral signatures of outliers.
- **Model Adaptation:** GNN architectures can be adapted to incorporate these spectral considerations, leading to models like the Beta Wavelet Graph Neural Network (BWGNN), which is specifically designed to handle the 'right-shift' effect inherent in outliers.

Theoretical Support: The 'right-shift' phenomenon has been rigorously proven on a Gaussian outlier model, validating its occurrence in a variety of graphs with synthetic or real-world outliers. This theoretical support underpins the design choices made in addressing the 'right-shift' through GNNs. By incorporating strategies to address the 'right-shift', GNNs can become more effective in detecting outliers within graph-structured data. The development of architectures like BWGNN represents a significant advancement in this direction, offering a theoretically justified and practically effective solution for the challenges posed by the 'right-shift' phenomenon..

4.3 Feature Extraction and Separation

BWGNN divides node characteristics into two groups; class characteristics (C) and neighboring characteristics (S). The class characteristics depict the qualities of the nodes while the neighboring characteristics convey the structural details. This division enables BWGNN to retain data, for outlier detection while minimizing the impact of signals, from diverse neighbors. By upholding the authenticity of node representations BWGNN guarantees that the model can accurately distinguish between irregular nodes thereby improving its ability to detect outliers[16].

4.4 Model Architecture

BWGNNs structure comprises elements :-

Spectral Filtering :- This involves applying filters to the graph Laplacian to focus on frequency components that are important, for detecting outliers.

Feature Aggregation :- It combines class features and neighboring features separately to ensure that the representations of nodes remain accurate.

Classification Module :- This component utilizes the aggregated features to categorize nodes as either normal or anomalous utilizing enhanced feature representations to enhance detection precision.

By integrating these elements BWGNN effectively tackles the challenges associated with SDS thereby enhancing its reliability and accuracy, in identifying outliers.

Chapter 5

EXPERIMENTS AND RESULTS

The experiments are designed to rigorously evaluate BWGNN's performance in detecting outliers within the Amazon and YelpChi datasets

5.1 Dataset

The Amazon dataset, characterized by its dense product- user interaction network, and the YelpChi dataset, noted for its segmented user-business review network, are pre-processed according to the methodology outlined in previous section.

Amazon
The Amazon dataset contains reviews of products sold on Amazon.com[17]. Each node represents a user or a product, and edges represent relationships such as "user rated product" or "product belongs to category."

Commonly used for tasks such as recommendation systems, where the goal is to predict user preferences or product ratings. Format: Typically provided as a collection of JSON files, where each file contains information about users, products, reviews, etc.

YelpChi
The YelpChi dataset consists of user reviews and social network information from Yelp[18]. Nodes represent users or businesses, and edges represent social connections or reviews.

Similar to the Amazon dataset, provided as JSON files containing information about users, businesses, reviews, etc.

5.2 Experimental Setup

The algorithms are implemented in Python and executed on a system having an Intel Core i7 processor with 16 GB RAM running on Windows 11.

5.3 Performance Metrics

Performance is assessed using precision, recall, F1 macro, GMean and the area under the receiver operating characteristic (ROC) curve (AUC-ROC), providing a comprehensive view of BWGNN's detection accuracy and reliability.

Precision: Precision is the ratio of correctly predicted positive observations to the total predicted positives. It measures the accuracy of the positive predictions. The formula for precision is:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{5.1}$$

Recall (Sensitivity): Recall is the ratio of correctly predicted positive observations to the all observations in actual class. It measures the ability of the model to find all the positive samples. The formula for recall is:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{5.2}$$

F1 Score (F1 Macro): The F1 score is the harmonic mean of precision and recall. It provides a balance between precision and recall, giving equal weight to both metrics. The formula for the F1 score is:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5.3}$$

For multiclass classification, the F1 macro score calculates the F1 score for each class and then averages them to give equal weight to each class.

G-Mean: The geometric mean (G-Mean) is a measure of classifier performance that balances sensitivity and specificity. It is the square root of the product of sensitivity and specificity. The formula for G-Mean is:

$$\text{G-Mean} = \sqrt{\text{Sensitivity} \times \text{Specificity}} \tag{5.4}$$

Area Under the ROC Curve (AUC-ROC): The ROC curve is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. The AUC-ROC measures the area under the ROC curve, which indicates the model's ability to distinguish between positive and negative classes[19]. AUC-ROC values range from 0 to 1, where a higher value indicates better performance.

5.4 Result Analysis

The application of BWGNN to the Amazon and YelpChi datasets yielded the results shown in Table 5.1.

Dataset $\longrightarrow$	**YelpChi**			**Amazon**		
Metrics $\longrightarrow$ **Models** $\downarrow$	**AUC**	**F1-macro**	**GMean**	**AUC**	**F1-macro**	**GMean**
SVM	70.37	70.77	0	90.51	90.71	0
GCN	56.51	51.31	45.51	86.67	60.54	76.38
PC-GNN	85.12	69.33	77.20	96.14	86.54	89.78
GDN	90.34	76.05	80.84	97.09	90.68	90.78
Proposed	**91.24**	**77.25**	**81.24**	**97.26**	**92.20**	**91.02**

Table 5.1: Results

5.4.1 Amazon Dataset:

BWGNN demonstrated exceptional precision and recall, significantly outperforming baseline models. The AUC-ROC score indicated a high degree of reliability in distinguishing between regular and outlier nodes within the dense network.

5.4.2 YelpChi Dataset:

Despite the dataset's segmented topology, BWGNN maintained its high performance, showcasing its adapt- ability to various graph structures. Its F1 score, in particular, underscored its balanced detection capability.

Chapter 6

CONCLUSION AND FUTURE SCOPE

6.1 Conclusion

This study introduced the Beta Wavelet Graph Neural Network (BWGNN), a novel approach to outlier detection in graph- structured data that leverages spectral graph theory and neural network architectures to effectively address the challenges posed by structural distribution shifts (SDS). By implementing BWGNN on the Amazon and YelpChi datasets, we demonstrated its superior performance in identifying outliers, outperforming existing methods including traditional GNNs and the Graph Decomposition Network (GDN).

6.2 Key Findings

BWGNN's integration of spectral graph analysis allows for a nuanced understanding of graph structures, enabling the detection of outliers with high precision and reliability across diverse datasets.The adaptability of BWGNN to various graph dynamics, as evidenced by its performance on the Amazon and YelpChi datasets, showcases its potential for broad application in detecting graph-based outliers. BWGNN's architecture and methodology present a significant advancement in addressing SDS, ensuring robust outlier detection even in evolving graph environments.

6.3 Future Directions

Exploring Additional Datasets Applying BWGNN to a wider range of graph-structured data can further validate its versa- tility and effectiveness in outlier detection.

Enhancing Model Efficiency Investigating methods to optimize BWGNN's computational efficiency could broaden its applicability, especially in real-time detection scenar- ios.

Integrating Advanced Spectral Techniques: Incorporating cutting-edge developments in spectral graph theory could enhance BWGNN's capability to uncover complex outlier patterns.

Cross-Domain Applications: Examining BWGNN's applica- bility across different domains, such as cybersecurity, finance, and social media analysis, could reveal new insights and use cases for graph-based outlier detection.

In conclusion, the Beta Wavelet Graph Neural Network (BWGNN) represents a significant step forward in the field of graph- based outlier detection. Its ability to seamlessly navigate the complexities of SDS and deliver high-performance results across varied datasets underscores the potential of spectral graph theory in advancing the detection of outliers in graph- structured data. We anticipate that further research and de- velopment will continue to uncover the full extent of BWGNN's applicability and impact in this evolving field

Bibliography

[1] N. Arun, S. Paul, and B. Srinivasan, "Improving robustness of graph neural networks to structure poisoning attacks," 2022.

[2] Y. You, T. Chen, Y. Shen, and Z. Wang, "Graph contrastive learning with augmentations," in *Advances in Neural Information Processing Systems*, 2020, pp. 5812–5823.

[3] J. You, R. Ying, and J. Leskovec, "Addressing structural distribution shifts for generalization in graph neural networks," in *International Conference on Learning Representations*, 2020.

[4] Author(s), "Optimal analysis of structures by concepts of symmetry and regularity," *Journal Name*, vol. Volume, no. Number, p. Pages, Year.

[5] J. Zhou, G. Cui, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun, "Graph neural networks: A review of methods and applications," *AI Open*, 2020.

[6] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," in *International Conference on Learning Representations*, 2018.

[7] W. L. Hamilton, R. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *Advances in Neural Information Processing Systems*, 2017, pp. 1024–1034.

[8] M. Defferrard, X. Bresson, and P. Vandergheynst, "Convolutional neural networks on graphs with fast localized spectral filtering," in *Advances in Neural Information Processing Systems*, 2016, pp. 3844–3852.

[9] W. Yu, W. Huang, T. Zhang, Y. Rong, H. Liu, and J. Huang, "Graph representation learning on heterogeneous networks via contextualized attention," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2020.

[10] M. Zhang, S. Cui, M. Neumann, Y. Chen, and T. Ma, "Graph distribution networks," in *International Conference on Machine Learning*, 2021.

[11] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2017.

[12] D. Chen, Y. Lin, W. Li, P. Li, and J. Zhou, "Can adversarially regularized graph autoencoder improve graph embedding?" in *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, 2019.

[13] J. Tang, J. Li, Z. Gao, and J. Li, "Rethinking graph neural networks for anomaly detection," in *International Conference on Machine Learning*, 2020.

[14] Y. Zhang, Y. Qian, D. Zhao, K. Ding, and H. Liu, "Graph adversarial training: Dynamically regularizing based on graph structure," in *Proceedings of the 2020 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2020.

[15] E. Rossi, B. P. Chamberlain, F. Frasca, D. Eynard, F. Monti, and M. Bronstein, "Temporal graph networks for deep learning on dynamic graphs," *arXiv preprint arXiv:2006.10637*, 2020.

[16] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[17] J. McAuley and J. Leskovec, "Hidden factors and hidden topics: understanding rating dimensions with review text," in *Proceedings of the 7th ACM conference on Recommender systems*, 2013, pp. 165–172.

[18] J. McAuley, C. Targett, and Q. Shi, "Inferring networks of substitutable and complementary products," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015, pp. 785–794.

[19] Author(s), *ICT Systems Security and Privacy Protection: 35th IFIP TC 11 International Conference, SEC 2020, Maribor, Slovenia, September 21–23, 2020, Proceedings*. Springer, 2020. [Online]. Available: https://dokumen.pub/ict-systems-security-and-privacy-protection-35th-ifip-tc-11-international-conference-sec-2020-maribor-slovenia-september-2123-2020-proceedings-1st-ed-9783030582005-9783030582012.html

ANNEXURE-IV

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-42

PLAGIARISM VERIFICATION

Title of the Thesis: Applications of Graph Neural Networks in Anomaly Detection

Total Pages: 24 Name of the Scholar: Rohit Saini Supervisor (s)

(1) Dr. Anurag Goel

(2) _____

(3) _____

Department: Computer Science & Engineering

This is to report that the above thesis was scanned for similarity detection. Process and outcome is given below:

Software used: Turnitin Similarity Index: 14%, Total Word Count: 5293

Date: _____

RohitSaini

Candidate's Signature **Signature of Supervisor(s)**

PAPER NAME

2K22_AFI_18_ROHIT_SAINI_MTech_Thesis.pdf

AUTHOR

ROHIT SAINI

WORD COUNT

5293 Words

CHARACTER COUNT

31579 Characters

PAGE COUNT

24 Pages

FILE SIZE

613.2KB

SUBMISSION DATE

Jun 10, 2024 9:15 PM GMT+5:30

REPORT DATE

Jun 10, 2024 9:16 PM GMT+5:30

14% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

- 9% Internet database
- Crossref database
- 8% Submitted Works database
- 7% Publications database
- Crossref Posted Content database

14% Overall Similarity

Top sources found in the following databases:

- 9% Internet database
- Crossref database
- 8% Submitted Works database
- 7% Publications database
- Crossref Posted Content database

TOP SOURCES

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	web.archive.org Internet	3%
2	Thu Uyen Do, Viet Cuong Ta. "Attention Pooling for Beta Wavelet Filter... Crossref	<1%
3	dokumen.pub Internet	<1%
4	Sheffield Hallam University on 2023-11-28 Submitted works	<1%
5	ijeast.com Internet	<1%
6	fastercapital.com Internet	<1%
7	"Graph Neural Networks: Foundations, Frontiers, and Applications", Spr... Crossref	<1%
8	arxiv.org Internet	<1%

9	link.umsl.edu Internet	<1%
10	Birkbeck College on 2018-09-11 Submitted works	<1%
11	University of Wollongong on 2021-02-22 Submitted works	<1%
12	University of Surrey on 2024-05-17 Submitted works	<1%
13	Xie, Shuisheng, Jundong Liu, and Charles D. Smith. "Riemannian Shape... Crossref	<1%
14	WeiDong Zhao, XiaoTong Liu. "Detection of E-Commerce Fraud Review... Crossref	<1%
15	adoc.pub Internet	<1%
16	Liverpool John Moores University on 2023-06-14 Submitted works	<1%
17	ebin.pub Internet	<1%
18	Aston University on 2024-05-23 Submitted works	<1%
19	assets.researchsquare.com Internet	<1%
20	fr.slideshare.net Internet	<1%

21	tudr.thapar.edu:8080 Internet	<1%
22	Coventry University on 2017-08-21 Submitted works	<1%
23	Hong Kong University of Science and Technology on 2021-11-17 Submitted works	<1%
24	dspace.daffodilvarsity.edu.bd:8080 Internet	<1%
25	stars.library.ucf.edu Internet	<1%
26	Texas A&M University, College Station on 2024-01-22 Submitted works	<1%
27	ikee.lib.auth.gr Internet	<1%
28	jes.ksu.edu.tr Internet	<1%
29	frontiersin.org Internet	<1%
30	Huang, Zexi. "Learning Representations for Information-Rich Graphs", ... Publication	<1%
31	Shouheng Li, Dongwoo Kim, Qing Wang. "Chapter 28 Beyond Low-Pass... Crossref	<1%
32	d197for5662m48.cloudfront.net Internet	<1%

33	export.arxiv.org Internet	<1%
34	Colorado State University, Global Campus on 2023-02-27 Submitted works	<1%
35	Heriot-Watt University on 2024-04-16 Submitted works	<1%
36	Hu, Chenhui. "Graph-Based Data Mining in Neuroimaging of Neurologic... Publication	<1%
37	Institute of Technology, Nirma University on 2014-12-16 Submitted works	<1%
38	Lingfei Ren, Ruimin Hu, Dengshi Li, Yang Liu, Junhang Wu, Yilong Zang,... Crossref	<1%
39	National University of Singapore on 2009-09-26 Submitted works	<1%
40	University of Queensland on 2024-04-08 Submitted works	<1%
41	University of Strathclyde on 2023-08-14 Submitted works	<1%
42	Vilnius Gediminas Technical University on 2024-04-21 Submitted works	<1%
43	dspace.dtu.ac.in:8080 Internet	<1%
44	rstudio-pubs-static.s3.amazonaws.com Internet	<1%

45	Imperial College of Science, Technology and Medicine on 2024-06-03 Submitted works	<1%
46	Meng, Lin. "Applying Graph Neural Networks to Applications: Brain Dis... Publication	<1%
47	Optimal Analysis of Structures by Concepts of Symmetry and Regularit... Crossref	<1%