GRAPH BASED ANOMALY DETECTION IN TEXTS

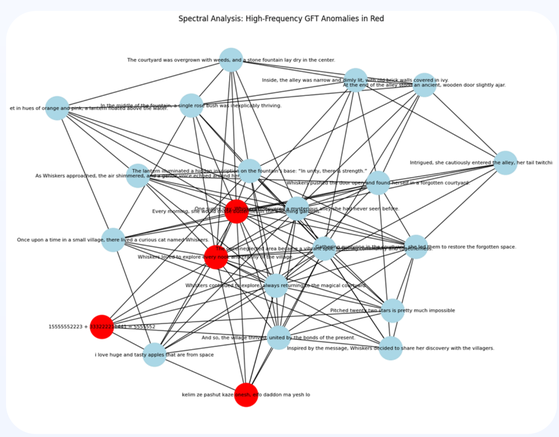
AUTHORS
Alexander Kovtonyuk

SUPERVISORS
Prof. Zeev Volkovich

AFFILIATES
Braude Academic College

INTRODUCTION

Anomaly detection is a data mining technique focused on identifying rare or unusual occurrences within datasets. These anomalies, or outliers, deviate significantly from the norm and can have crucial implications across various domains. Unlike traditional outlier detection methods that analyze data as independent points in multi-dimensional space, anomaly detection in graphs leverages the inherent relationships and dependencies between data objects.



OBJECTIVE

Develop a Graph-Based Anomaly Detection Framework.

- Combine semantic embeddings (BERT) with graph representations for text analysis
- Use spectral decomposition to identify high-frequency signals as anomalies.
- Test with synthetic anomalies in real-world corpora.

METHODOLOGY

Input Text Data

Collect raw text
Preprocess the text
tokenize sentences
filter and create chunks.

Embedding Generation

Use a pre-trained BERT model to generate embeddings for each text chunk

Graph Construction

Build a k-NN Graph
Nodes: Text chunks
Edges: Weighted by similarity scores

Spectral Analysis

Compute the normalized Laplacian of the graph.
Identify high frequency components using eigenvalues and eigenvectors.

Anomaly Detection

Detect nodes with the highest amplitudes in the high-frequency signal

ANALYSIS

We validated our approach across multiple datasets, testing various k values in the k-NN graph to assess their impact on anomaly detection. Lower K values highlighted anomalies but failed to differentiate them by high-frequency amplitudes. As K increased, the graph captured more contextual relationships, improving detection accuracy and balancing precision with recall.

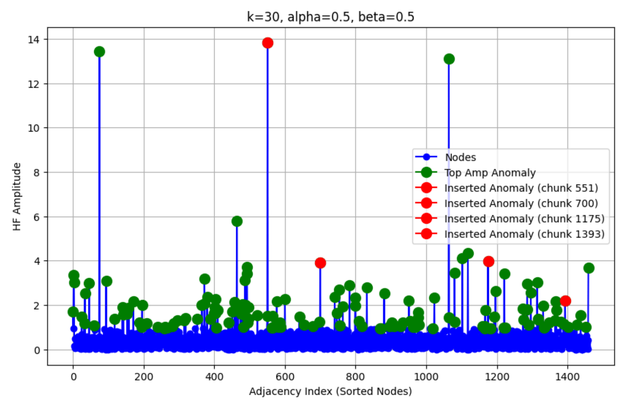


Figure 1, K=30, Dataset project gutenberg “MOBY-DICK or, THE WHALE”

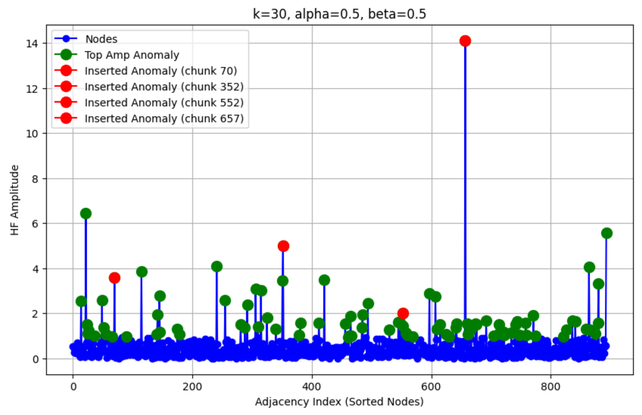


Figure 2, K=30, Dataset project gutenberg “Pride and Prejudice”

RESULTS

We can see in the plot graphs that the pipeline is able to recognize anomalies in different datasets of classical literature.
The anomalies are highlighted in red, the green nodes represent the top 10% nodes with the highest amplitude.

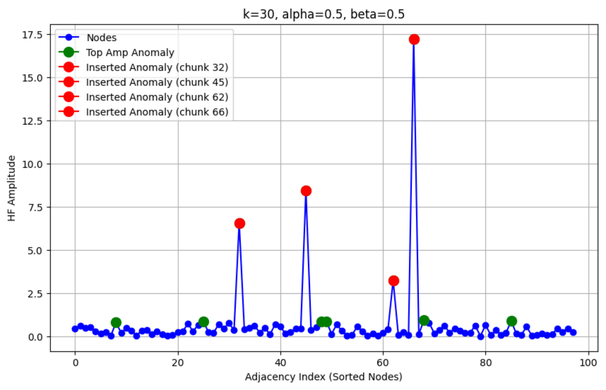


Figure 3, K=30, Dataset “The Greek New Testament”

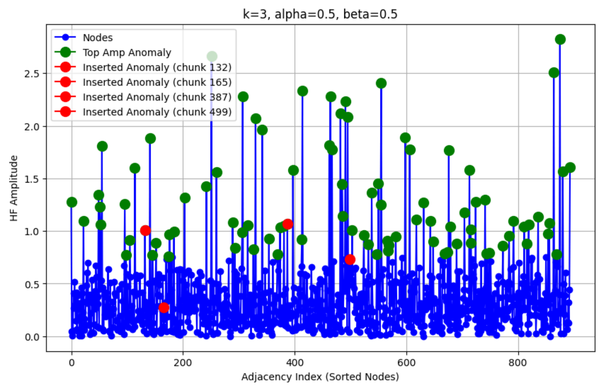


Figure 4, K=3, Dataset project gutenberg “Pride and Prejudice”

CONCLUSIONS

Our approach successfully identifies anomalies by detecting high-frequency components in the graph representation of text chunks. Through validation on different k-NN graphs, we observed that larger K-values improved connectivity but required more pronounced anomalies to stand out. While the method identifies anomalies, it also produces false positives, indicating a need for improved embeddings or alternative graph construction strategies.

FUTURE WORK

One key area for improvement is enhancing graph construction process, We aim to explore graphs or weighted edges based on semantic similarity. Additionally, alternative embeddings, such as Sentence Transformers, GPT-based models, or domain-specific representations, could improve the model’s ability to distinguish anomalies. We also plan to investigate advanced spectral techniques, including wavelet transforms and spectral clustering, to refine anomaly detection further.

KEY IMPROVEMENTS

- **Improved Graph Construction** – Testing different graphs construction methods or weighted edge enhancements.
- **Optimized Graph Fourier Transform** – Exploring different spectral methods to improve anomaly separation.
- **Fine-tuning Embeddings** – Using domain-adapted or fine-tuned transformer models for more precise vector representations