ANALYZING ARTICLE'S CITATIONS USING - ANOMALY DETECTION IN DYNAMIC GRAPHS

Authors: Elroei Seadia, Dvir Bublil

24-2-R-2

Supervisor: Prof. Zeev Volkovich

Advisor Dr. Renata Avros

Software Engineering
Department Braude

Introduction

Accurate detection of anomalous citation patterns in academic literature is crucial for maintaining research integrity, identifying potential misconduct, and uncovering emerging research trends. Citation networks can be effectively represented as dynamic graphs, where articles are nodes and citations are edges. This representation allows us to capture the temporal evolution of these networks as a series of snapshots, reflecting the changing relationships between articles over time. To address this challenge, we adapt the Anomaly Detection in Dynamic Graphs via Transformer (TADDY) framework, which excels in identifying anomalies in dynamic graphs By leveraging TADDY and implementing subsequent analyses, we aim to not only detect anomalous citation patterns but also to understand and extract the relationships between these anomalies and the associated articles. This enables us to examine trends in anomalous article behavior and visualize the spread of anomalous citations across the years covered by the dataset.

Methodology

The research process involves collecting and preprocessing PubMed data by cleaning and deduplicating it, then transforming it into dynamic graphs where articles are nodes and citations are edges. These graphs are analyzed as sequential snapshots to track evolving connections. Using spatial-temporal encoding and a transformer-based model, anomalies are detected by learning citation patterns. The model, trained on both normal and anomalous data, is evaluated using the ROC-AUC metric to ensure accurate detection. Finally, results are visually presented to facilitate in-depth analysis and trend tracking over time.

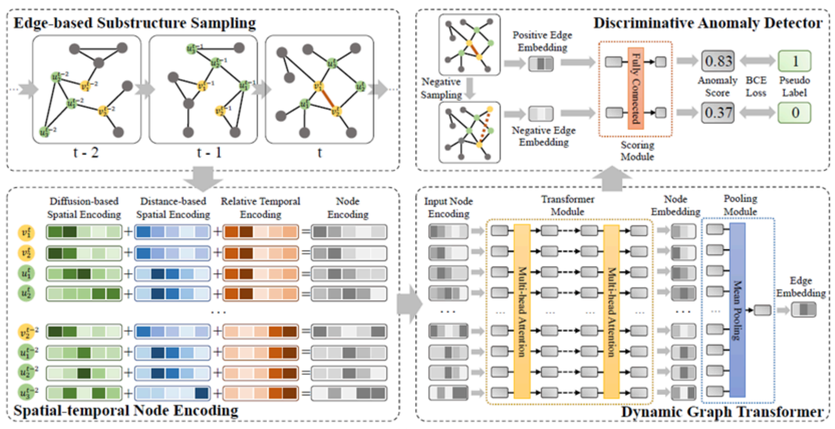

Figure 1: Key components: substructure sampling, spatial-temporal encoding, a dynamic graph transformer, and an anomaly detection module.

Results

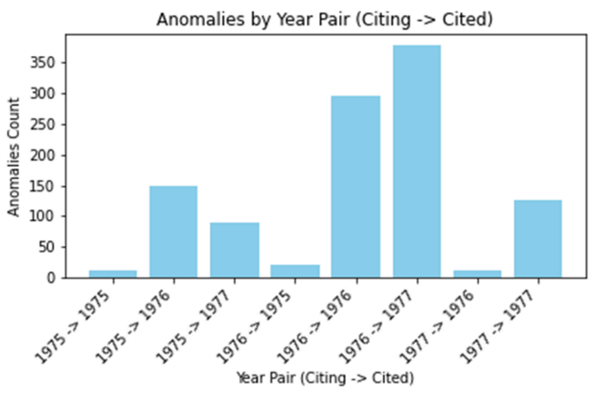

Fig. 16 shows the distribution of anomalies by year pairs (Citing -› Cited), with the highest concentration in 1977 -› 1977, likely due to the large volume of research published that year.

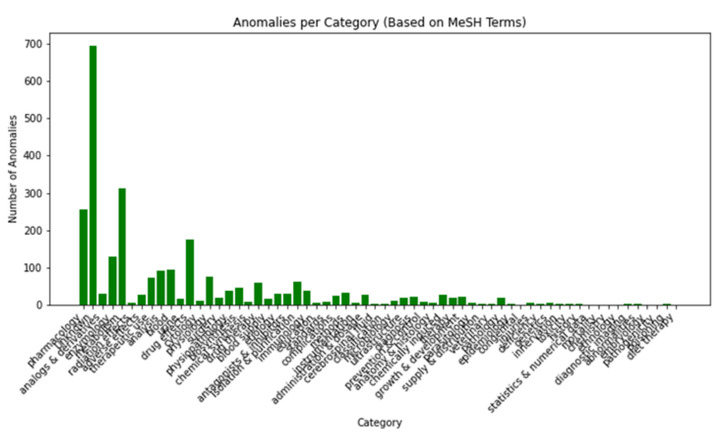

Fig. 17.Illustrates the distribution of anomalies across various categories based on MeSH terms. As we can see a lot of articles MeSH terms was unknown because of lake in the dataset. (for example: Drug effect, genetics, pharmacology)

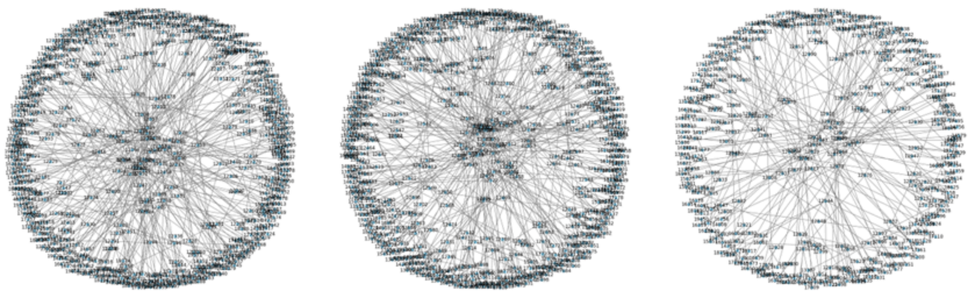

Fig.4 Three distinct states of the citation graph showcasing evolving structure.

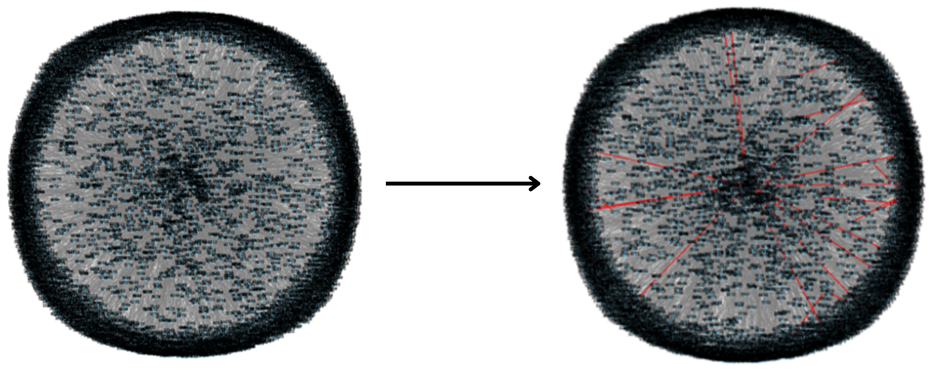

Fig. 5. Show the snapshots graph structure built from citations and articles, the red edges in the right shows the same snapshot with the anomalous edges.

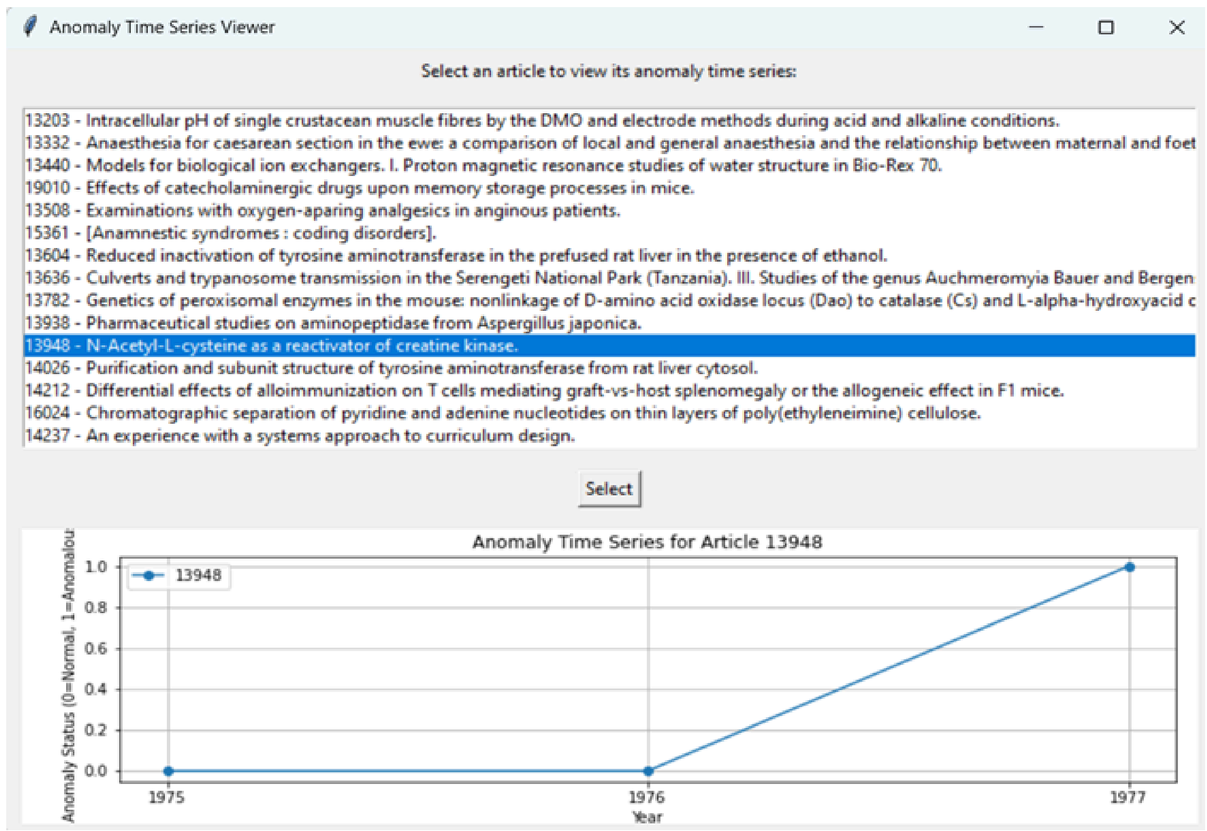

Fig. 18.Illustrates selection of anomalous article. We can see during 1975 – 1976 he was normal and detect as anomalous in 1977

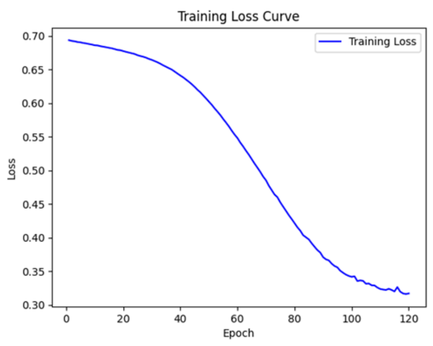

Fig. 2. The training loss curve over 120 epochs.

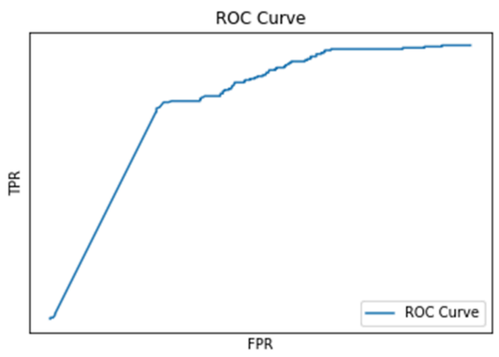

Fig. 3 ROC curve shows the model's performance

Conclusion

The research demonstrates the effectiveness of the TADDY framework in detecting anomalous citation patterns, achieving high accuracy in distinguishing anomalies over time. The visual tools provide valuable insights for tracking citation trends and articles.