Detecting and Evaluating Anomalously Cited Papers Over Time using Anomaly Detection in Dynamic Graphs via Transformers

Noah Soskha and Israel Shushan

Software Engineering Department Braude Academic College

Introduction

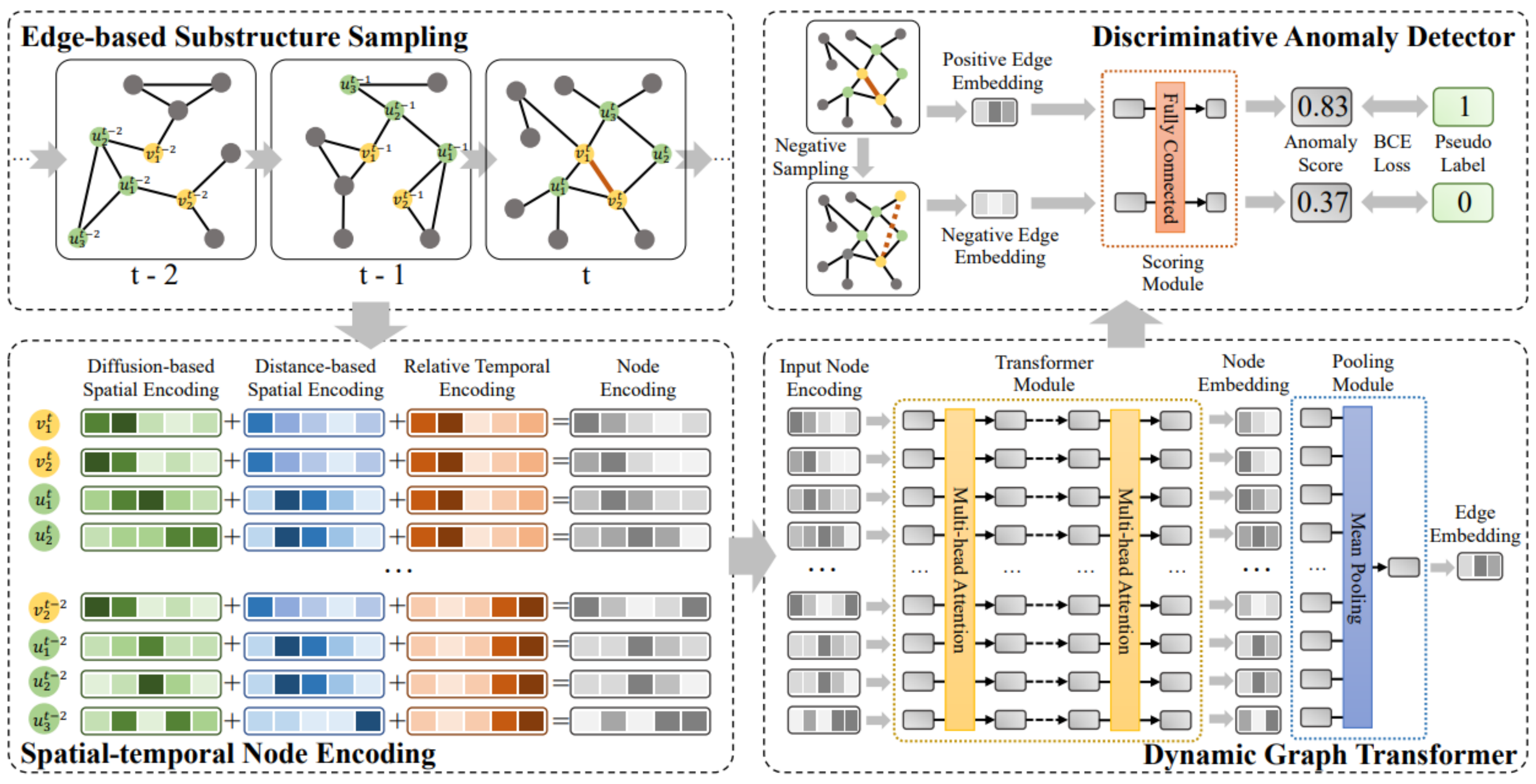
Academic citation networks are fundamental in tracking the flow of knowledge across scholarly works. However, citation manipulation and irregular patterns pose threats to the integrity of these networks. Traditional methods struggle to detect temporal and complex citation anomalies, especially when citation patterns evolve over time. As citation manipulation, self-citation clusters, and unexplained citation spikes distort research evaluations, the need for more effective anomaly detection becomes crucial. This project proposes a solution that utilizes advanced machine learning methods, specifically the TADDY framework, to detect anomalous citation behaviors dynamically, improving research evaluation and enhancing academic integrity.

Methodology

This project applies the **Transformer-based Anomaly Detection in Dynamic Graphs (TADDY)** framework to identify and track anomalous citation behaviors across multiple timeframes. It integrates both **spatial** and **temporal** graph features to analyze academic citation networks. This approach consists of multiple stages, including graph preparation, contextual node encoding, anomaly detection, and evaluating.

Main Steps of the algorithm:

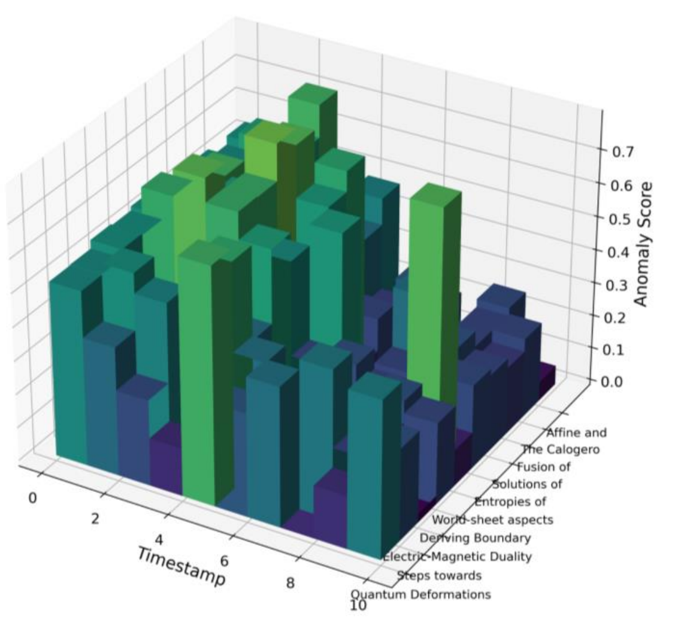
- **Data Input**: Citation graph data is represented as nodes (papers) and edges (citations) with temporal and structural attributes.
- **Node Encoding:** A combination of diffusion-based, distance-based, and temporal encodings capture both structural and time-dependent relationships.
- **Transformer Model:** The Dynamic Graph Transformer processes the node encodings, generating edge embeddings that represent citation relationships.
- **Anomaly Detection:** A discriminative anomaly detector, implemented as a fully connected neural network, takes edge embeddings (both normal and pseudo-anomalous) as input and computes an anomaly score for each edge, minimizing a loss function that distinguishes between them to enable effective learning over multiple iterations.
- **Track Changes:** Track temporal patterns of anomalous edges through visualizations including histograms, standard distributions, and temporal anomaly evolution metric of the highest-scoring anomalous edges. After that we report the papers that these citations (edges) stem from.

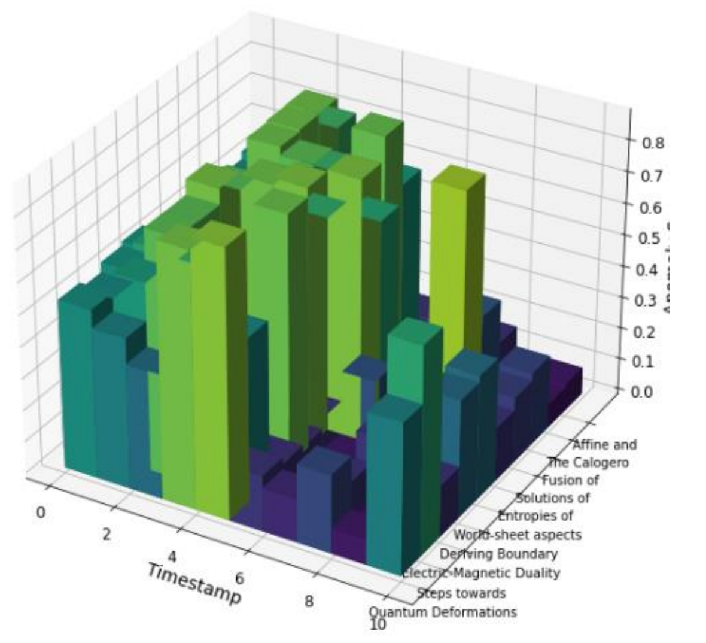


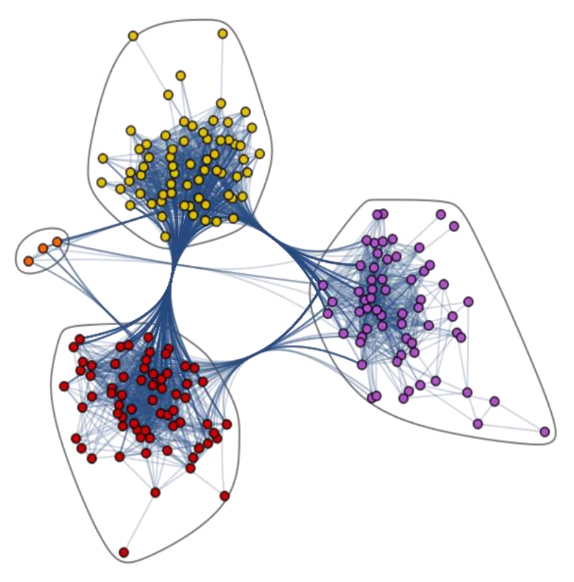
Challenges Review

- **Data Preprocessing**: Large integer node identifiers required optimization for efficient processing. The solution was to remap node identifiers into a compressed range.
- **Dependency Management**: Due to complex interdependencies and outdated libraries, we implemented version control through Anaconda to ensure reproducibility and resolve compatibility issues.
- **Dataset**: Identifying suitable datasets proved challenging, as established citation networks like CORA and PUBMED lacked the necessary metadata for our analysis.

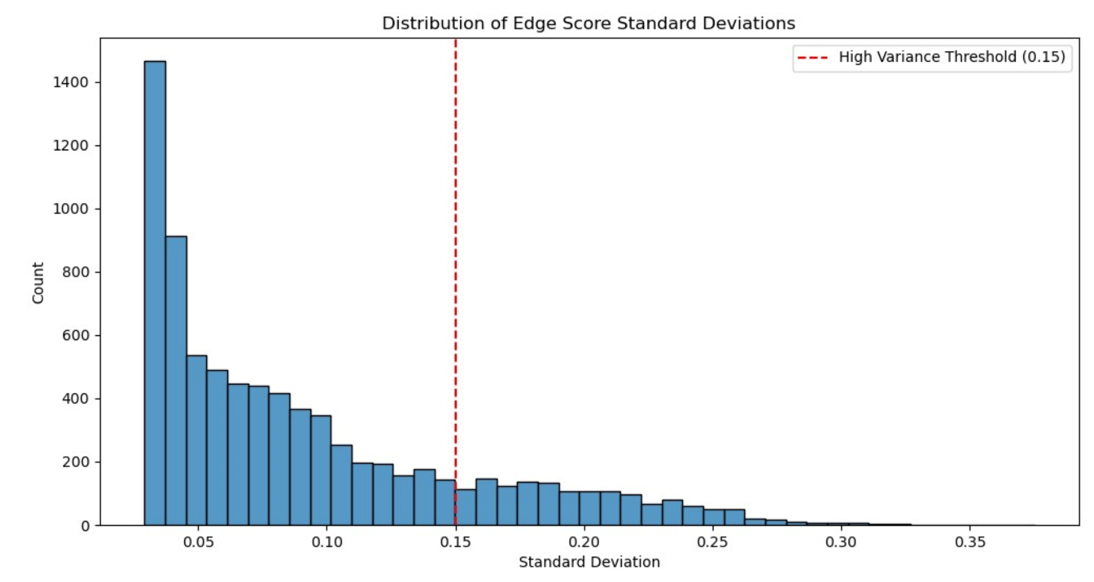
Network Dynamics

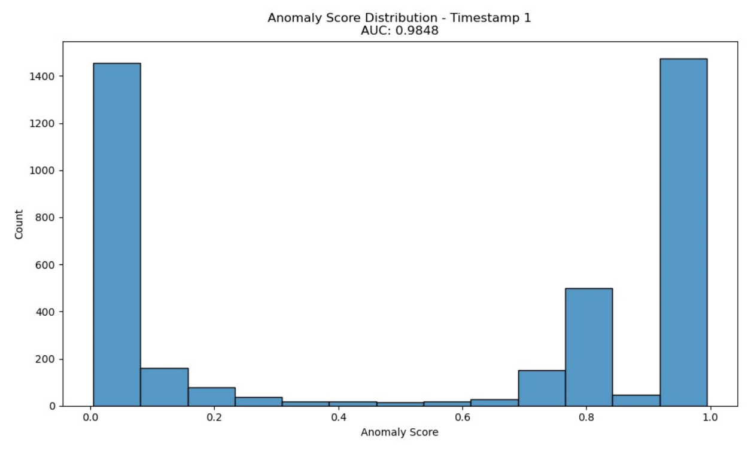


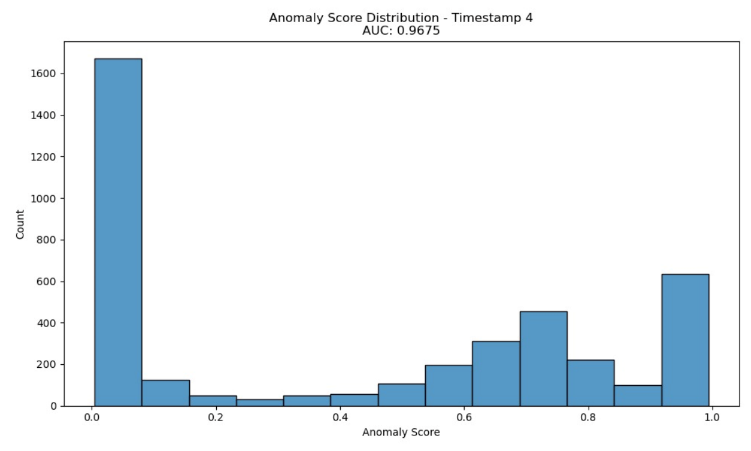


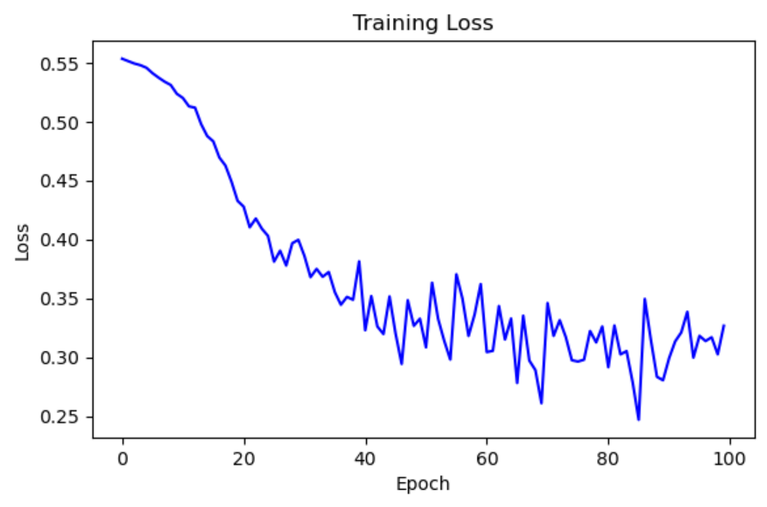


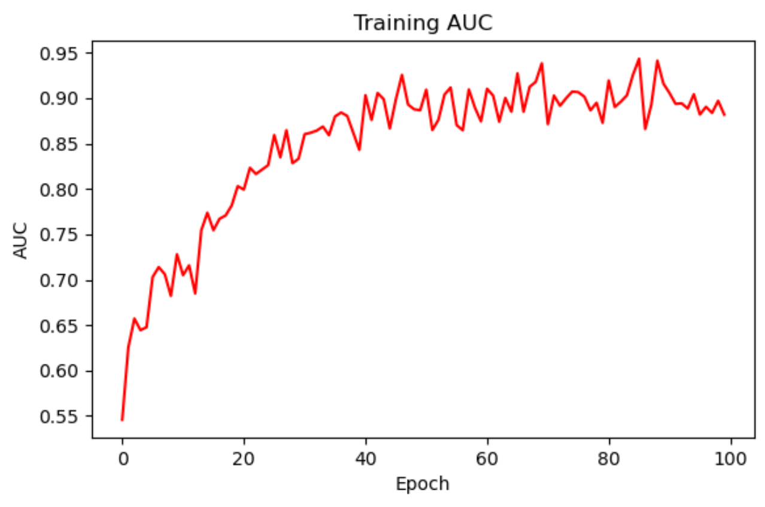
Results

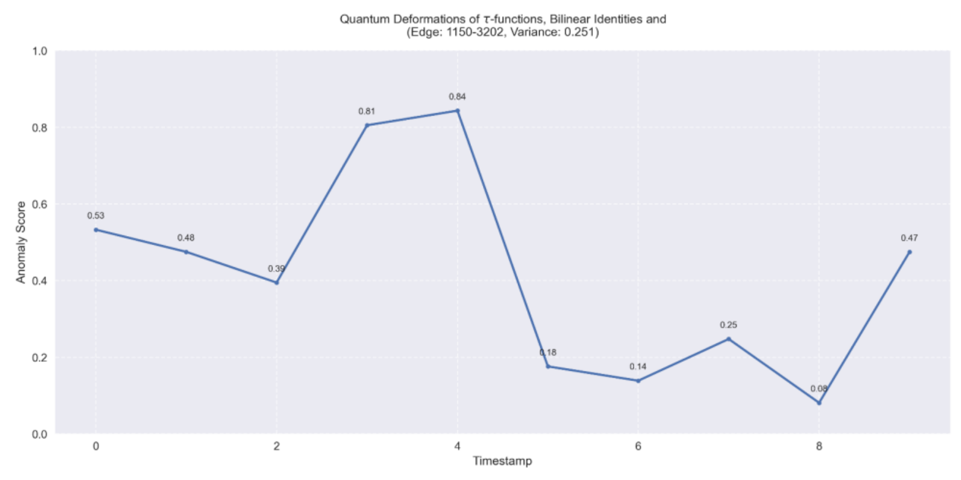


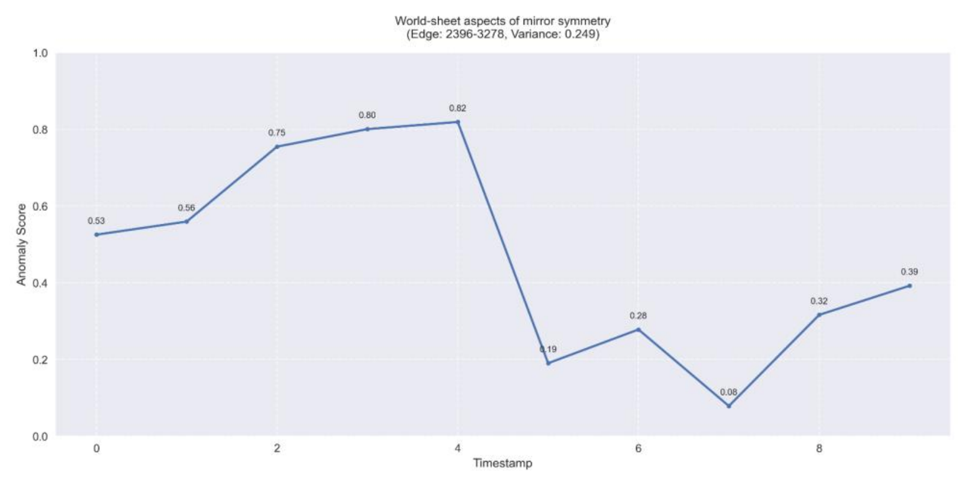












- **Edge Score Standard Deviation Distribution** shows citation behaviors that have low variance, with only a few edges surpassing the high variance threshold (0.15), indicating potential anomalies.
- **Anomaly Score Histogram:** The histograms illustrate distinct patterns across timestamps. At Timestamp 6, the distribution shows a sharp peak near 0.6, while Timestamp 13 reveals a bimodal distribution with major clusters around 0.0 and 0.6-0.8, indicating a more polarized classification of anomalies.
- **Training Loss and AUC** demonstrates an effective model learning, with loss decreasing from 0.55 to approximately 0.30 and AUC improving significantly from 0.55 to stabilize around 0.90. The AUC curve shows stability after epoch 40, despite minor fluctuations.
- **Temporal Anomaly Pattern:** The figures track two papers identified with high variance scores (≈0.25) in their citation patterns. The graphs demonstrate how their anomaly scores evolved across timestamps, showing significant fluctuations ranging from peaks of 0.80-0.84 to lows of 0.08.

Conclusion

This project explores the dynamic nature of anomalies in citation networks. Using a transformer-based model, it offers an advanced perspective on how citation patterns evolve across space and time. By incorporating pseudo-anomalous edges, the model enables effective training even without labeled anomalies, showcasing its practicality for real-world citation networks.
The findings indicate that most papers maintain stable citation patterns, reflecting their typical impact on scientific research. However, a subset of articles exhibits notable variations in citation behavior, signaling changes in their influence within the scientific community. These articles merit closer examination and thoughtful analysis.

Acknowledgments

Thank to our supervisors, Prof. Zeev Volkovich and Dr. Renata Avros, for their invaluable guidance and support throughout this research.