# 

**Software Engineering Department Braude**

**Academic College**

**Capstone Project Phase A - 61998**

# Detecting and Evaluating Anomalously Cited Papers Over Time using anomaly detection in dynamic graphs via transformer

24-2-R-7

Table of Contents

[Abstract 2](#_Toc176351517)

[1. Introduction 2](#_Toc176351518)

[The Importance of Detecting Citation Anomalies 3](#_Toc176351519)

[2. Background 3](#_Toc176351520)

[2.1.1 Graph Theory and Dynamic Graphs 3](#_Toc176351521)

[2.1.2 Subgraphs and Graph Diffusion 3](#_Toc176351522)

[2.1.3 Node and Edge Embeddings 3](#_Toc176351523)

[2.1.4 Discriminative Knowledge 4](#_Toc176351524)

[2.1.5 Spatial Information 4](#_Toc176351525)

[2.1.6 Temporal Information 4](#_Toc176351526)

[2.1.7 Citation Network 4](#_Toc176351527)

[2.1.8 Neural Networks 4](#_Toc176351528)

[2.1.9 Dynamic Graphs 5](#_Toc176351529)

[2.1.10 Graph Transformer Models 5](#_Toc176351530)

[2.1.11 Attention Mechanisms 5](#_Toc176351531)

[3. Mathematical Background 5](#_Toc176351532)

[3.1 Graph Theory 5](#_Toc176351533)

[3.1.1 Dynamic Graphs 5](#_Toc176351534)

[3.1.2 Subgraph and Substructure 6](#_Toc176351535)

[3.1.3 Graph Diffusion 6](#_Toc176351536)

[3.1.4 Personalized PageRank (PPR) 7](#_Toc176351537)

[3.1.5 Heat Kernel 7](#_Toc176351538)

[3.2 Node Encoding 7](#_Toc176351539)

[3.2.1 Diffusion-based Spatial Encoding 7](#_Toc176351540)

[3.2.2 Distance-based Spatial Encoding 8](#_Toc176351541)

[3.2.3 Relative Temporal Encoding 8](#_Toc176351542)

[3.3 Transformer Model 9](#_Toc176351543)

[3.3.1 Attention Mechanism 9](#_Toc176351544)

[3.3.2 Multi-head Attention 9](#_Toc176351545)

[3.4 Node and Edge Embedding 10](#_Toc176351546)

[3.4.1 Node Embedding 10](#_Toc176351547)

[3.4.2 Edge Embedding 10](#_Toc176351548)

[3.5 Anomaly Detection 11](#_Toc176351549)

[3.5.1 Anomaly Score Calculation 11](#_Toc176351550)

[3.5.2 Binary Cross-Entropy Loss 11](#_Toc176351551)

[3.6 Negative Sampling 12](#_Toc176351552)

[4. Research Process 13](#_Toc176351553)

[4.1 Model Overview 13](#_Toc176351554)

[4.2 Datasets 13](#_Toc176351555)

[4.3 Model Architecture 13](#_Toc176351556)

[4.3.1 Initialization 13](#_Toc176351557)

[4.3.2 Embedding 13](#_Toc176351558)

[4.3.3 Testing Stage 13](#_Toc176351559)

[4.3.4 Model Evaluation 13](#_Toc176351560)

[4.4 Processing and Evaluation 13](#_Toc176351561)

[4.5 Anomaly Detection Techniques 14](#_Toc176351562)

# Abstract

In academic research, the integrity of citations is essential to the validity and reliability of academic work. However, the increasing complexity and volume of citations have led to challenges ensuring that all references are relevant and contribute meaningfully to the research. The main problem lies in detecting anomalous cited papers that were cited irrelevantly. To address this challenge, we propose a systematic approach that uses anomaly detection techniques. Specifically, we utilize TADDY, a transformer-based anomaly detection framework, to analyze citation patterns within dynamic graph where papers are nodes and edges are citations. By applying TADDY, we use its results to create an "anomaly score" for each paper in our dataset. The use of TADDY anomaly detection techniques ensures that our approach can handle the evolving nature of academic citations. Our results provide a reliable mechanism for enhancing the integrity of academic references.

**Keywords:** Citation Anomalies, Citation Network, Dynamic Graphs, Transformer Models, Anomaly Detection

# Introduction

Citation networks are used in measuring the scholarly impact of academic papers. These networks consist of nodes and edges, where papers are represented as nodes and the citations between them as edges. Anomalous citations—those that deviate from standard citation patterns—can indicate irregularities or manipulations in how papers reference each other. Examples include citations that do not logically support the citing paper's claims, those aimed at artificially boosting citation metrics, or those that are inappropriate given the paper's topic.

Our study focuses on detecting and analyzing these anomalies to determine how they evolve over time. As new papers are published, what may initially appear as an anomalous citation can eventually be seen as justified, remain anomalous, or become non-anomalous as the dataset evolves. This could indicate whether the citation was indeed anomalous.

### The Importance of Detecting Citation Anomalies

Reliability of citation practices plays a pivotal role in maintaining the authenticity and relevance of scholarly communication. Citation anomalies, ranging from irrelevant references to manipulative citation practices, threaten this integrity. These anomalies could manifest as citations that do not support the claims they are purported to bolster or are used strategically to inflate citation metrics unfairly. Such practices not only skew the actual impact of research but also mislead readers and researchers, potentially developing a skewed academic environment where the true contribution of research might be overshadowed by manipulated citation figures.

# Background

### Graph Theory and Dynamic Graphs

The study of dynamic graphs, also known as temporal graphs, is essential for understanding how relationships and structures evolve over time within datasets. This concept is particularly relevant in academic citation networks where the relationships between papers (nodes) and their citations (edges) can change, reflecting the evolving nature of scholarly discourse.

### Subgraphs and Graph Diffusion

In the context of detecting citation anomalies, understanding subgraphs—subsets of a graph’s nodes and edges that form a smaller graph—is vital. This understanding aids in focusing on local citation patterns that might not be evident when considering the larger network. Additionally, graph diffusion processes, such as Personalized PageRank, help in understanding the broader structure of citation networks by disseminating information across the graph, thereby depicting the influence and connectivity of nodes over time.

### Node and Edge Embeddings

To effectively detect anomalies in citation networks, it is crucial to have a solid method of representing the nodes (papers) and edges (citations). Embedding techniques, which transform nodes and edges into a vector space of lower dimensionality, facilitate this by capturing the essential characteristics of nodes and their links within the graph. These embeddings serve as a foundational input for subsequent anomaly detection processes.

### Discriminative Knowledge

Discriminative knowledge encompasses insights that effectively distinguish between various categories within datasets by highlighting features that are uniquely associated with specific classes. This type of knowledge is helpful in classification tasks, where the objective is to identify key attributes that differentiate one group from another. In the context of our research, discriminative knowledge is leveraged to discern patterns of citation that deviate from normative academic practices, aiding in the detection anomalous citations within scholarly documents.

### Spatial Information

Spatial information holds data related to the position, size, and relationships among physical locations in space. In our research, spatial information relates to how nodes positioned relative to each other and how these positions affect their interactions or connections.

### Temporal Information

within a dataset, Temporal information captures changes and developments over time. In dynamic graphs, temporal information will help in understanding how relationships and interactions evolve. This type of data can reveal trends, patterns, and anomalies. this information allows us to analyze temporal patterns in citation practices, helping to detect shifts that may indicate emerging trends or irregular citation activities.

### Citation Network

Citation networks are a type of directed graph where nodes represent academic papers, and directed edges signify citations from one paper to another. By examining the structure and connectivity of these networks, researchers can infer the influence of specific works and detect trends and patterns in academic disciplines.

citation networks are analyzed to uncover potentially anomalous citation behaviors. we identify unusual patterns of citations that may not align with typical scholarly practices.

### Neural Networks

Artificial neural network (ANN) models its architecture after the human brain to function within the domain of machine learning. It stands as a valuable tool in artificial intelligence, equipped to enable computers to learn from data and make informed decisions.

Neural networks are organized into layers containing nodes, each designed to perform similarly to a neuron in the human brain. These include an input layer, several hidden layers, and an output layer. Nodes are interconnected, each bearing weights and thresholds. When the output from a node surpasses its threshold, the node activates and transmits data to subsequent layers in the network.

### Dynamic Graphs

Dynamic graphs are data structures that evolve over time by adding or removing nodes and edges. They are useful in scenarios where relationships between entities are not static but fluctuate with time, such as citation networks. Dynamic graphs enable the analysis of how structures grow and transform. Dynamic Graphs will help in analyzing how citation links between academic papers develop and deviate from expected patterns over time.

### Graph Transformer Models

Graph Transformer Models are a class of deep learning models designed to manage graph-structured data by applying the principles of transformer architectures, these models use attention mechanisms to dynamically weigh the relationships between nodes in a graph, allowing them to learn contextual relationships and dependencies with high precision. This method enables processing and analysis of relational and structural information in graphs. These models use attention mechanisms to weigh the importance of nodes and their connections, learning complex patterns and dependencies within the graph data. This approach is particularly valuable in tasks that require processing and interpreting the complex relationships between papers and authors, detecting anomalies in citation behaviors.

### Attention Mechanisms

Attention mechanisms are a component of modern neural network architectures, originally developed to improve models like language translation. These mechanisms allow the Graph Transformer Model to dynamically focus on various parts of the input data, assigning varying levels of 'attention' to each part based on its relevance to the task at hand. Which will enable the model to make more nuanced decisions by considering context and dependencies more effectively.

# Mathematical Background

## Graph Theory

### Dynamic Graphs

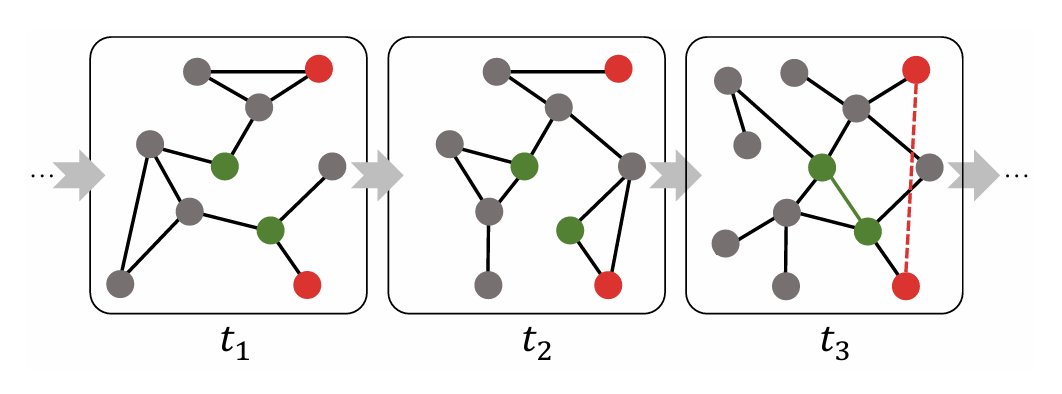
A dynamic graph, is represented as a series of snapshots , where each snapshot contains a set of nodes and edges at time t. The changes in the graph structure over time capture the dynamic interactions and relationships between entities.

Fig. 1. [Dynamic Anomaly Detection Example](https://arxiv.org/pdf/2106.09876v1) - A toy example to illustrate how the coupled information affects the detection of edges’ legality. The three graphs are a fragment from a dynamic graph stream at a sequential timeline . The solid green line represents a normal edge at , while the red dash line indicates an anomalous edge. We highlight the corresponding nodes in the previous timestamps with colors.

### Subgraph and Substructure

A subgraph is a subset of a graph's nodes and edges that forms a graph. Substructures in anomaly detection refer to local subgraphs centered around target edges or nodes. These substructures help detect anomalies by focusing on local patterns that may not be apparent in the global graph structure.

This approach allows for efficient computation and localized analysis, which is crucial in large-scale dynamic graphs.

תמונה שמכילה תרשים, קו

התיאור נוצר באופן אוטומטי

Fig. 2. [Edge-based Substructure Sampling](https://arxiv.org/pdf/2106.09876v1)

In edge-based substructure sampling, the target nodes (in yellow) and contextual nodes (in green) from multiple timestamps are sampled to construct the substructure node set, where neighboring node number k and time window size are both set to be 3.

### Graph Diffusion

Graph diffusion processes are used to capture the global structure of a graph by spreading information from each node to its neighbors. Two common types of graph diffusion are Personalized PageRank (PPR) and the heat kernel.

### Personalized PageRank (PPR)

Personalized PageRank (PPR) computes a ranking score for each node based on its connectivity to a specific node. The diffusion matrix for PPR is defined as:

where α is the teleport probability, is the identity matrix, D is the diagonal degree matrix, and A is the adjacency matrix.

This formulation allows for capturing both local and global information in the graph by balancing between direct connections and global reachability.

### Heat Kernel

The heat kernel diffusion captures the spread of heat (information) over time in a graph. The diffusion matrix for the heat kernel is defined as:

where is the diffusion time, and exp represents the matrix exponential.

The heat kernel is effective in modeling processes that involve gradual propagation, such as heat distribution or information flow in networks.

## Node Encoding

Node encoding is essential for capturing the structural and temporal information of nodes in a dynamic graph. Various encoding strategies are used to create informative representations of nodes.

### Diffusion-based Spatial Encoding

Diffusion-based spatial encoding uses the diffusion matrix to capture the global structure of nodes in the graph. The rank-based encoding is computed using the following formula:

whereis the index function, rank is the ranking function, and linear is a learnable linear mapping.

This encoding captures the influence and importance of a node within the graph.

### Distance-based Spatial Encoding

Distance-based spatial encoding represents the local structure around the target edge by computing the minimum distance to the target nodes. The encoding is defined as:

where is the distance function, is the minimum function, and linear is a learnable linear mapping.

This encoding effectively captures the local neighborhood information, which is important for identifying local anomalies.

### Relative Temporal Encoding

Relative temporal encoding captures the temporal information of nodes by computing the difference between the occurrence time of the target edge and the current timestamp. The encoding is given by:

where is the absolute difference in time, and linear is a learnable linear mapping.

This encoding is crucial for capturing temporal dynamics and changes over time, enabling the detection of temporal anomalies.

תמונה שמכילה טקסט, צילום מסך, עיצוב

התיאור נוצר באופן אוטומטי

Fig. 3. [Spatial-temporal Node Encoding](https://arxiv.org/pdf/2106.09876v1) - Three types of encoding are computed for each node and further fused into the node encoding.

## Transformer Model

Transformers are a powerful class of neural networks originally designed for sequence data, but they have been adapted for graph data due to their ability to capture complex dependencies.

### Attention Mechanism

The attention mechanism allows the model to focus on various parts of the input sequence, which is crucial for capturing relationships in graph data. The scaled dot-product attention is defined as:

where **Q**ueries, **K**eys and **V**alues are input matrices, and is the dimension of the keys.

This mechanism enables the model to weigh the importance of different nodes and edges dynamically.

### Multi-head Attention

Multi-head attention extends the attention mechanism by applying it multiple times in parallel, with each attention head focusing on various parts of the input. It is defined as:

where each head is an instance of the attention mechanism, and is a learnable parameter matrix.

This allows the model to capture multiple aspects of the data simultaneously, enhancing its ability to learn complex patterns.

תמונה שמכילה טקסט, צילום מסך, מספר, מקביל

התיאור נוצר באופן אוטומטי

Fig. 4. [Dynamic Graph Transformer](https://arxiv.org/pdf/2106.09876v1) Three types of encoding are computed for each node and further fused into the node encoding.

## Node and Edge Embedding

Embedding techniques are used to represent nodes and edges in a lower-dimensional space, capturing their essential characteristics.

### Node Embedding

Node embeddings are computed using transformer layers, which aggregate information from the node's neighbors. The dimension of the embeddings is controlled by a parameter

These embeddings capture the structural and relational properties of nodes in the graph.

### Edge Embedding

Edge embeddings are obtained by pooling the node embeddings within a substructure. The mean pooling operation is used to aggregate the embeddings:

where is the matrix of node embeddings, and is the number of nodes in the substructure.

This aggregation captures the overall relationship and interaction between nodes, which is crucial for edge-level anomaly detection.

## Anomaly Detection

Anomaly detection involves identifying data points that deviate significantly from the expected patterns. In this context, the focus is on detecting anomalous edges in dynamic graphs.

### Anomaly Score Calculation

Anomaly scores are computed using a fully connected neural network with a Sigmoid activation function. This neural network serves as the discriminative anomaly detector, which distinguishes between normal and anomalous edges based on their embeddings. The anomaly score for an edge is given by:

where is the edge embedding, is a weight matrix, and is a bias term.

This score indicates the likelihood of an edge being anomalous based on its embedding.

### Binary Cross-Entropy Loss

The discriminative anomaly detector is trained using a binary cross-entropy loss function, which measures the discrepancy between the predicted and actual labels. The loss function is defined as:

Where and are the positive and negative edges, respectively.

This loss function guides the model to distinguish between normal and anomalous edges.

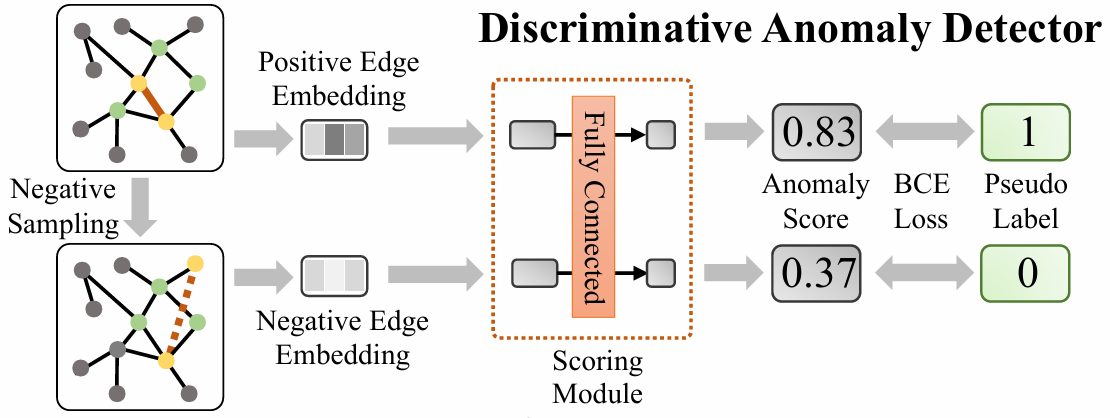


Fig. 5. [discriminative anomaly detector](https://arxiv.org/pdf/2106.09876v1) - the negative edges are acquired by negative sampling. The scoring module computes the anomaly scores for positive and negative edges. The whole framework is trained with a binary cross-entropy loss in an end-to-end manner.

## Negative Sampling

Negative sampling is used to generate negative samples for training the anomaly detector. The strategy ensures that negative samples do not appear in the positive edge set, providing a ~~robust~~ training signal.

This technique is crucial for handling imbalanced datasets and improving the model's ability to discriminate between normal and anomalous edges.

# Research Process

## Model Overview

This section provides an overview of the TADDY algorithm's application in our research, focusing on detecting and analyzing citation anomalies within dynamic citation networks. The aspect of our approach includes extending the algorithm's capability to capture temporal dynamics more comprehensively by implementing a 3D visualization of anomaly scores over time.

## Datasets

We utilize the CORA dataset, recognized for its rich citation networks and compatible for demonstrating the TADDY algorithm's efficiency in identifying anomalies within scholarly communications. The dataset's complex network of citations provides a robust basis for testing our enhanced anomaly detection methods.

## Model Architecture

### Initialization

The initialization phase involves configuring the TADDY algorithm with predefined parameters suitable for processing the CORA dataset, including data loading and feature normalization setups.

### Embedding

During this stage, citation data is transformed into a vector space to accurately capture and represent the relationships and attributes of papers and their citations, essential for the subsequent anomaly detection process.

### Testing Stage

Post-embedding, the model evaluates the transformed data against established citation patterns to identify potential deviations indicating anomalies.

### Model Evaluation

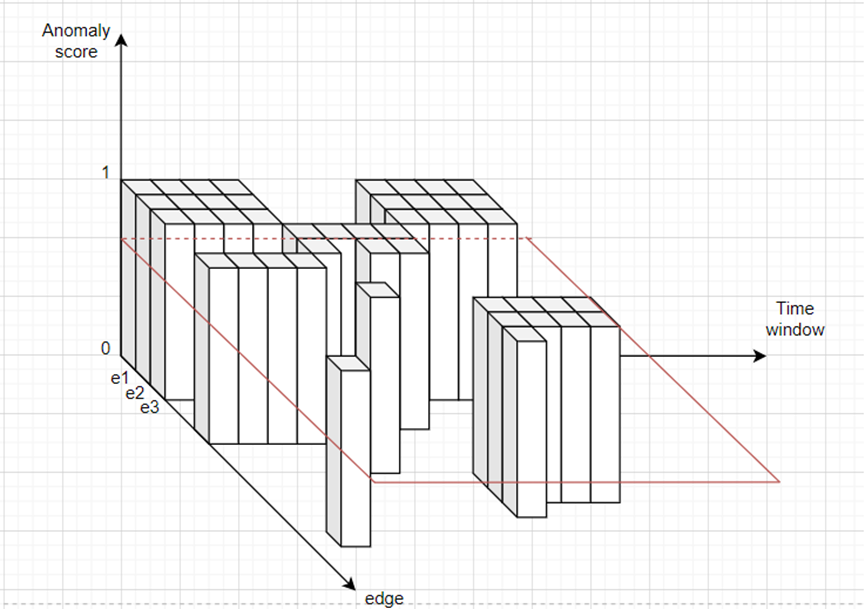
The model's effectiveness is gauged by comparing its anomaly detection outputs against a manually reviewed set of citations. Visualization techniques are employed to depict these comparisons, aiding in intuitive understanding and analysis.

## Processing and Evaluation

In this final phase, the data processed through the TADDY algorithm is meticulously analyzed. This comprehensive evaluation involves several testing and validation steps to ensure the robustness and accuracy of the detected anomalies, which are crucial for refining the model's capabilities.

## Anomaly Detection Techniques

This section introduces an extension to the conventional application of the TADDY algorithm, emphasizing our development of a three-dimensional temporal analysis method. This enhances our ability to analyze the progression of citation anomalies over time within academic papers.

**** Fig. 6. This is an example of our score function for a specific paper.

The coordinates of this function are: (Time window, edge from the graph, anomaly score).

Where the function is:

The red plane represents the threshold height that has been predefined.

To deepen our analysis, we adapt the TADDY algorithm to output a vector of anomalous edges not just at a single timestamp but across multiple time windows. This allows us to construct a dynamic 3D function for each paper, where the axes represent time windows, edges, and corresponding anomaly scores.

For each time window evaluated, the output vector from TADDY updates this 3D function for the anomaly score associated with each targeted paper. Each edge ​ at its respective time window ​ adjusts the function's value, marking it distinctly to denote an anomaly. This iterative update provides a robust framework for observing the citation's evolution.

By intersecting each function with a horizontal plane at a set threshold height, we then calculate the cumulative sum of the function's area above this threshold. This approach not only captures but also quantifies the shifts in citation behavior over time, highlighting specific periods where anomalous activities are most pronounced.