mathematics

MDPI

Article

Spotting Suspicious Academic Citations Using Self-Learning Graph Transformers

Renata Avros, Mor Ben Haim, Almog Madar, Elena Ravve and Zeev Volkovich *

Software Engineering Department, Braude College of Engineering, Karmiel 2161002, Israel; ravos@braude.ac.il (R.A.); mor.ben.haim@e.braude.ac.il (M.B.H.); almog.madar@e.braude.ac.il (A.M.); cselena@braude.ac.il (E.R.)
* Correspondence: vlvolkov@braude.ac.il

Abstract: The study introduces a novel approach to identify potential citation manipulation within academic papers. This method utilizes perturbations of a deep embedding model, integrating Graph-Masked Autoencoders to merge textual information with evidence of graph connectivity. Consequently, it yields a more intricate model of citation distribution. By training a deep network with partial data and reconstructing masked connections, the approach capitalizes on the inherent characteristics of central connections amidst network perturbations. It demonstrates its ability to pinpoint trustworthy citations within the analyzed dataset through comprehensive quantitative evaluations. Additionally, it raises concerns regarding the reliability of specific references, which may be subject to manipulation.

Keywords: graph-masked autoencoders; manipulated citations; network perturbation

MSC: 37M05

Citation: Avros, R.; Haim, M.B.; Madar, A.; Ravve, E.; Volkovich, Z. Spotting Suspicious Academic Citations Using Self-Learning Graph Transformers. *Mathematics* **2024**, *12*, 814. https://doi.org/10.3390/math12060814

Academic Editor: Dmitriy V. Ivanov

Received: 20 December 2023
Revised: 3 March 2024
Accepted: 7 March 2024
Published: 10 March 2024

1. Introduction

Manipulated citations involve intentionally including references in academic works to gain somewhat biased advantages compared to their authentic academic merit. Rather than supporting the author's arguments or providing relevant background information, the primary aim is to artificially boost the citation count of the cited author, their works, a specific journal, and so forth. So, several key aspects are associated with manipulated citations commonly used to enhance researchers' perceived impact and prestige and artificially inflate journals and metrics.

The detrimental effects of this practice reach far, jeopardizing the cornerstone of academic discourse—precision, impartiality, and scientific credibility. Although researchers acknowledge the uneven value of citations and attempt to address it by differentiating and assigning weights based on type, most studies only focus on this specific approach, neglecting broader considerations. Prabha [1] sheds light on the vastness of the problem by revealing that over two-thirds of references in a paper are deemed unnecessary, providing further evidence of the widespread presence of dubious citations.

The most prevalent forms of citation manipulation include excessive self-citation, where an author cites their irrelevant work, and coercive citation, in which reviewers or editors pressure authors to cite specific work, including the author's own or publications in particular journals. Additionally, practices such as citation rings and ghost citations warrant consideration in this context.

The manipulation of academic citations poses a serious threat to the scientific community. This common unethical practice erodes trust in research, distorts the research landscape by inflating citation counts for personal gain, and misallocates resources. Measures like journal policies, reviewer training, researcher transparency, and metric improvement are used to

combat this. Comprehending and addressing manipulated citations is pivotal for upholding the integrity of academic research, safeguarding its credibility, and ensuring reliability.

Numerous surveys were conducted to investigate the practices of manipulating reference lists. It is worth highlighting that this issue deserves special attention [2–4]. Several authors acknowledged that these citations, similar to rumors, differ from the touchstone in their context and comportment compared to normal, regular references.

While conventional techniques such as manual inspection and basic statistical analyses were utilized, they come with limitations in capturing intricate patterns and subtle manipulations. In recent years, network-based approaches have emerged as promising methods for identifying and comprehending citation manipulation. The intricate nature of graph data, characterized by irregular structures and relational dependencies, poses a challenge for conventional anomaly detection techniques.

By harnessing the inherent structure and connections within citation networks, network-based approaches can reveal hidden relationships and anomalies indicative of potential citation manipulation. These methods transcend individual paper analysis, delving into the broader network dynamics to foster a more comprehensive understanding of manipulation patterns. Many research papers explored this avenue, including [5–8]. While primarily focused on deep learning methods for identifying rumors and fake news within networks, their insights and methodologies are valuable in developing effective network-based approaches for detecting citation manipulation in academic research.

In contrast to traditional methods, anomaly detection approaches leveraging graph learning possess the capability to simultaneously preserve both node attributes and network structures throughout the learning process, offering a more suitable methodology for addressing the complexities associated with graph data. By utilizing the structure and connections within the citation network, network-based approaches can uncover hidden relationships and abnormalities indicative of potential citation manipulation. These methods extend beyond individual papers, delving into broader network dynamics, thereby providing a more comprehensive understanding of manipulation patterns. Papers [7,9] present research of this kind.

This study builds upon the general approach presented in [10], focusing on the concept of characterizing connections within a citation graph by analyzing their behavior under network perturbations. In other words, the conjecture is that genuine relationships within the network are more resistant to disruptions. The approach is suggested using the Node2Vec method [11], which provides a graph embedding handling random walks within a graph. The research is predicated on the idea that manipulated or fraudulent citations manifest anomalies within a citation network. These anomalies make them vulnerable to appropriate network perturbations, resulting in instability and detectability. The hypothesis is grounded in the belief that manipulated citations, strategically inserted to enhance the impact or credibility of specific publications artificially, deviate from the natural patterns and structures inherent in the citation network. Consequently, when exposed to network perturbations such as removing specific nodes or edges, manipulated citations are more likely to exhibit inconsistencies that distinguish them from genuine citations.

Even with certain limitations in its applicability, the offered implementation of the general concept leads to demonstrably acceptable results that exhibit a high degree of concordance with empirical data. However, one of this approach's disadvantages is, time and again, ignoring the papers' textual component and resting upon just inner network connectivity.

This article presents a method that harnesses the structural connections within a citation network and also the textual similarities between articles using Graph-Masked Autoencoders (GMAEs) suggested in [12]. Such an approach extension makes it possible to clarify and generalize the previously obtained results.

Employing graph transformers (e.g., [12,13]) demonstrates promising performance in learning graph representations, which appears natural. Unlike traditional transformers, which process sequential data like text, graph transformers handle data represented as graphs, taking into account the inner graph structure and the features of the nodes to learn

informative representations of nodes by considering both their attributes and the attributes and relationships of their connected nodes. Like traditional transformers, they use attention mechanisms to focus on relevant parts of the graph, allowing them to capture long-range dependencies and complex relationships.

Despite their effectiveness, applying deep transformers in real-world scenarios poses challenges. Training them from scratch requires significant resources, and their memory consumption grows quadratically with the number of nodes, further hindering their practical implementation. GMAEs are introduced in [13] as a self-supervised transformer-based model for acquiring graph representations to tackle these challenges. GMAEs utilize a masking mechanism and an asymmetric encoder–decoder architecture to address the aforementioned limitations. Comprehensively, GMAEs take partially masked graphs as input and reconstruct the features of the masked nodes. The encoder–decoder architecture is deliberately designed with asymmetry, utilizing a deep transformer for encoding and a shallow transformer for decoding. In conjunction with the masking mechanism, this specific design makes GMAEs notably more memory-efficient than conventional transformers.

The approach suggested in this paper involves a network perturbation type that systematically removes a fixed set of nodes and reconstructs their features by leveraging GMAE trained on the remaining nodes. The iterative procedure includes substantial link prediction based on the omitted node features' recovery. Assessing the stability of citation reconstruction amid these node-masking perturbations has the potential to unveil abnormal citations. These anomalies could serve as cues for potential manipulation or fraudulent behavior within the citation network.

The proposed approach primarily targets typical or "standard" citation patterns, and it is essential to recognize its inherent limitations. Consequently, it may introduce inaccuracies, mainly when applied to multidisciplinary articles or instances known as "sleeping beauties". In this context, "sleeping beauty" refers to a research article that initially attracts minimal attention and remains unnoticed for an extended period post-publication. However, it experiences a significant surge in recognition and citations after a period of dormancy. Several factors contribute to this phenomenon, including groundbreaking developments that make previously overlooked research highly relevant or the discovery of the paper by other researchers who recognize its significance.

Moreover, it is important to note the citation, typically not acknowledged, of significantly ancient works authored by figures such as Newton, Archimedes, and others. Another challenge arises when dealing with multidisciplinary papers that suggest using several datasets from different research areas. One potential approach is to treat each citation against the appropriate collection. However, alternative setups can also be considered.

The rest of the manuscript is structured as follows: Section 2 outlines the pertinent mathematical foundations for the study. Section 3 introduces the proposed model designed for detecting citation manipulation. The experimental study assessing the efficacy of the model is outlined in Section 4. The paper concludes in Section 5 by summarizing key findings and discussing their implications.

2. Preliminaries

A graph, denoted as $G = (V, E)$, consists of a set of vertices (nodes) denoted by V and a set of edges denoted by E. If a graph has N_V nodes and N_E edges, an $N_V \times N_V$ adjacency matrix A denotes the connections between nodes. Each element in A is assigned a value of either 1 or 0, indicating the presence or absence of an edge between the corresponding nodes. Occasionally, graphs may include additional information, such as node features X_V (with dimension d_V) and edge features X_E (with dimension d_E).

The Graphormer architecture, presented in [14], addresses the challenges transformers face in adapting to graph structures. These problems include capturing relational information and managing the complexities of large graphs. The paper introduces a novel network design to overcome the mentioned limitations while leveraging the strengths of transformers.

The suggested model incorporates positional embeddings that inject structural information into the transformer architecture. This is achieved by encoding node centrality (in-degree and out-degree) into the transformer. Furthermore, the model captures pairwise node relationships through shortest path distances and integrates these distances as biases in the attention mechanism. The utilization of edge features enhances the overall ability of the Graphormer model to handle graph representations effectively.

The paper [13] suggests the utilization of Graph-Masked Autoencoders (GMAEs) with Graphormer as the foundational model. Specifically, both the encoder and decoder components are designed as graph transformers, inheriting their architecture from the Graphormer model. This implies that the transformer-based structure introduced by Graphormer serves as the backbone for both encoding and decoding processes within the context of Graph-Masked Autoencoders. The paper explores the application of Graphormer's transformer design in the GMAE framework, aiming to enhance the capabilities of graph autoencoding tasks.

Drawing inspiration from existing work, the paper considers Masked Language Modeling, a technique for learning representations from partially masked data. This approach aims to reduce training complexity and memory footprint while providing an opportunity to evaluate the stability of the general model.

The GMAE study employs a research framework that integrates the following components:

- Masking Mechanism: GMAEs take partially masked graphs as input, where a predetermined number of nodes is intentionally masked. This selective masking reduces the amount of information the model needs to process simultaneously, increasing memory efficiency during training.
- Asymmetric Encoder–Decoder Architecture: The GMAE model adopts an asymmetric architecture, employing a deep transformer encoder to extract rich representations from the unmasked nodes in the graph. On the other hand, the decoder consists of a shallower transformer network. The role of the decoder is to reconstruct the features of the masked nodes based on the encoded information obtained from the encoder. This design choice may contribute to a more effective and efficient information flow within the model.
- Self-Supervised Learning: GMAEs are trained using a self-supervised learning approach. In this context, the model is tasked with predicting the features of the masked nodes from the remaining information in the graph. This self-supervised learning paradigm is advantageous as it eliminates the dependency on labeled data, which are often scarce or expensive to obtain in real-world scenarios. The model learns to capture meaningful representations and relationships within the graph by leveraging the data's intrinsic structure.

GMAE aims to overcome challenges regarding memory efficiency, information reconstruction, and the accessibility of labeled data in graph-based tasks by incorporating these technical details.

In general, the forward propagation of GMAE comprises the following four steps:

1. Randomly mask nodes in the input graph.
2. Feed the non-masked nodes into the encoder and obtain their embeddings.
3. Use a shared learnable mask token to represent the embeddings of the masked nodes and insert them into the encoder output.
4. Feed the embedding matrix with inserted mask tokens into the decoder to reconstruct the features of the masked nodes.

A random subset of nodes is sequentially masked throughout the training phase for an input graph. The encoder, crucial to this process, is intentionally unaware of these masked nodes. It exclusively processes the features of the nodes that remain observable and subsequently generates embeddings for each of these observed nodes.

Node positional embeddings are enriched with centrality and spatial and edge encodings similar to Graphormer, necessitating knowledge of node degrees, all-pairs shortest paths, and (optionally) edge features.

3. Proposed Approach

This section introduces our approach to identifying anomalous citations in academic networks that potentially signify manipulation or fraud. The fundamental assumption, akin to one discussed in a previously published article [10], is that manipulated citations, strategically inserted to enhance the impact of particular publications, deviate from the natural structure of the network. These manipulated citations are expected to show inconsistencies and stand out when subjected to perturbations. We hypothesize that analyzing citation stability under perturbations, such as node removal, can reveal these deviations and identify suspicious citations. We aim to unveil potentially fraudulent behavior within the network by investigating anomalies and analyzing deviations from expected patterns.

The perturbations introduced in the considered citation network bear some resemblance to those discussed earlier in perturbation analyses of models involving artificial modifications to network structures. Specifically, within the context of the citation network, these perturbations entail the random removal of nodes corresponding to papers. These deliberate alterations simulate various scenarios or conditions to assess the citation network's robustness, stability, integrity, and individual links. Such perturbations serve as a mechanism to unveil vulnerabilities or weaknesses in a network. They increase the likelihood of anomalies or manipulated elements manifesting abnormal behavior or standing out amidst genuine components.

In the subsequent phase, link prediction using embeddings is carried out. Following the acquisition of embeddings, the similarity or proximity between pairs of nodes is quantified using diverse similarity metrics such as cosine similarity, Euclidean distance, or graph-based measures like mutual neighbors or the Jaccard coefficient.

The citation graph under consideration is treated as undirected, emphasizing the connections between papers rather than the specific directionality of citations. This focus on connectivity enables a comprehensive analysis of the network's structure and patterns by capturing the relationships and dependencies between papers, regardless of their citing or cited status.

We propose including two additional parameters to enhance the link prediction process: a similarity measure (S) and a threshold value (Tr). The similarity measure gauges the similarity between pairs of nodes, while the threshold value serves as the cutoff point for determining whether pairs are considered "connected" or not. Specifically, if the similarity score between two nodes surpasses the threshold (Tr), they are deemed connected, while pairs with a similarity score below the threshold are considered disconnected. This approach facilitates a nuanced and customizable evaluation of link predictions based on the defined similarity measure and threshold.

In broad terms, an adapted approach to evaluating the reliability of citations involves the following steps:

1. Load a graph $G = (V, E)$, including additional information containing node features XV (with dimension dV).
2. Repeat $Niter$ times:
    a. Randomly mask a fraction Fr of nodes in the input graph.
    b. Feed the non-masked nodes into the encoder and obtain their embeddings.
    c. Use a shared learnable mask token to represent the embeddings of the masked nodes and insert them into the encoder output.
    d. Calculate the similarity score for all pairs of the masked nodes using the measure S.
    e. Reconstruct the network of the omitted masked nodes by identifying potential links with similarity scores that meet or surpass the threshold (Tr).
3. For each connection, count how many times it is rebuilt throughout the iterations.

As was previously mentioned, in the GMAE model, the encoder and decoder exhibit asymmetry. The encoder is a deep graph transformer, while the decoder is a shallow graph transformer. The experiments in [13] found that employing an encoder with 16 layers and a decoder with 2 layers yields state-of-the-art performance in most cases. This design choice

results in an expressive encoder, optimizing performance, and it simultaneously conserves computational resources. Despite the depth of the encoder, the input feature matrix's size is reduced due to the masking mechanism. In contrast, a conventional end-to-end graph transformer employs a deep transformer similar to our encoder but utilizes a full feature matrix as input, leading to considerable memory consumption.

Conversely, the decoder input is an embedding matrix of full size, potentially implying large memory consumption. However, given the shallowness of the decoder, the computational load remains relatively small. This strategic design balances expressive power and computational efficiency in the encoder and decoder components of the GMAE model.

An ego-graph refers to a subset of a network focused on a specific node. This subset encompasses all nodes directly linked to the ego and the edges connecting them. Picture it as a zoomed-in perspective, providing a detailed view of a node's immediate neighborhood within the broader network. In GMAE, some nodes are "masked" during training, meaning their features are hidden. This makes it challenging to learn their connections to other nodes directly. Ego-graphs come in handy here because

- Focused analysis: By focusing on the ego-graph of a masked node, the model can concentrate its resources on reconstructing the missing connections for that specific node.
- Similarity-based reconstruction: GMAE utilizes similarity scores between nodes to infer potential connections. The ego-graph provides a smaller, more manageable context for comparing the similarity of neighboring nodes to the masked node, making the reconstruction process more efficient.
- Threshold-based filtering: The model can set a threshold for the similarity score. Only edges with similarity scores exceeding this threshold are considered potential connections for the masked node. This helps avoid reconstructing spurious connections based on weak similarities.

Focusing on ego graphs allows for a more accurate reconstruction of connections for masked nodes than analyzing the entire network. By limiting the scope of analysis, ego graphs reduce the computational burden of the reconstruction process. As the network size increases, ego graphs remain manageable, making the GMAE model scalable to large datasets.

So, during each training epoch (iteration), the model utilizes one ego graph for each node in the training data. In our application, each ego-graph has a depth of 1, working only with immediate neighbors (1-hop), randomly choosing maximum L_2 neighbors. As this value approaches infinity, the selection process becomes essentially unrestricted, and all neighbors are included based on their actual presence in the network.

Like GMAE, our method follows the approach outlined in Graphormer [14] to extract ego-graphs to train the model. Specifically, we leverage the neighbor sampler, as introduced in GraphSAGE [15], to create subgraphs by randomly sampling a designated number of nodes from the neighborhood of the target node. GraphSAGE operates by iterative sampling and aggregating information from a node's immediate neighbors. This involves the neighbor sampler selecting a subset of neighbors for each node during each iteration, reducing computational complexity and memory usage compared to processing the entire neighborhood. Our approach involves sampling from the immediate surroundings of each vertex. This method ensures that the model focuses on the relevant information within the node's neighborhood, contributing to the efficiency and effectiveness of the training process. Each ego-graph captures the central node's direct connections (one hop), excluding further indirect connections.

4. Experiments and Results

GMAE explores a variety of settings for its encoder layers, ranging from 1 to 30 while maintaining a constant of two layers for the decoder. Additionally, the mask ratio, determining the percentage of nodes subjected to masking, is adjusted between 0.7 and 0.8 with a step size of 0.1. The hidden dimensions are set at 64 for each layer, and each transformer layer incorporates eight attention heads. A linear decay learning rate scheduler is applied to enhance the training process, starting with a warm-up stage of 40,000 steps and gradually reducing the

learning rate to a final value of 1×10^{-9} after a maximum of 400,000 training steps. The peak learning rate is defined as 1×10^{-4}.

Our implementation incorporates the EarlyStopping Hooks callback from the PyTorch Lightning library to address a specific task to detect early signs of process stabilization. This callback is employed to halt training when a monitored metric ceases to improve. It is initialized with four parameters. The first parameter, 'metric', is set to 'train_loss'. The second parameter, 'mode', is set to 'min', signifying that training will conclude when the monitored metric stops decreasing. The third parameter, 'patience', is set to '500', indicating the number of training epochs with no improvement, after which training will be terminated. In this scenario, training will stop if the monitored metric shows no improvement for five hundred training epochs. The fourth parameter, 'check_on_train_epoch_end', is set to 'True'. When true, the callback assesses whether to stop training after each training epoch.

The current experiments adopt a dual focus, aiming to comprehend the scrutinized citation network structure and the method's effectiveness in identifying potentially irrelevant citations. To address this, each test is conducted twice, involving two versions of a given dataset. The initial version is the dataset in its original form, while the second version introduces noise by adding random connections, comprising 20% of the original connections in the source.

4.1. Cora Dataset

The Cora dataset, accessible at (URL: https://graphsandnetworks.com/the-cora-dataset/ (accessed on 6 March 2024)), stands as a well-established and extensively utilized resource in the realms of machine learning and natural language processing. Its principal focus lies in the exploration of citation networks. Comprising a collection of diverse scientific research papers, predominantly from computer science, the dataset spans various subfields such as machine learning, artificial intelligence, databases, and information retrieval. Each paper within the dataset is represented by a bag-of-words feature vector, indicating the presence or absence of specific words in the document. Furthermore, the Cora dataset furnishes details on citation links between papers, facilitating the examination of citation patterns and exploring techniques for analyzing citation networks.

With 2708 publications across seven categories and 5429 citation links, the Cora dataset offers a comprehensive and well-organized resource for studying scientific literature. Each publication is further characterized by a binary word vector of 1433 elements, where each element indicates the presence or absence of a specific word from the provided dictionary.

A partial visualization of the CORA dataset is given in Figure 1.

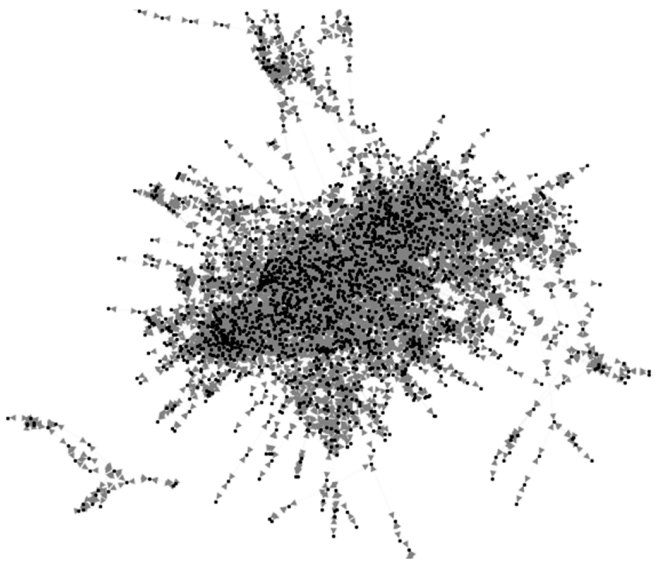

Figure 1. Partial visualization of the CORA dataset.

Experiments are performed with the following set of parameters:

- d = 64 (Embedding dimension).
- *N_encoder_layers* = 4 (Number of encoder layers).
- *N_decoder_layers* = 2 (Number of decoder layers).
- L_2 = 5/2000 (Numbers of neighbors in ego-graphs).
- *N_iter* = 50. (Number of epochs in the training process).
- *Fr*= 30%. (The fraction of the omitted nodes).
- *S*—the cosine similarity.
- *Tr—0.9/0.95*. (The link prediction threshold).
- *Peak_lr*—1×10^{-4}–*end_lr* 1×10^{-9}. (Learning rate).
- *Batch size* = 64.
- *Dropout rate* = 0.5.
- *Num_heads* = 8.
- *N_iter* = 1000. (The maximal number of iterations in GMAE training).

Cosine similarity functions as a metric for assessing the similarity between two vectors within a vector space by determining the angle's cosine. This yields a numerical representation indicating the degree of similarity. The scale of cosine similarity ranges from -1 to 1, where 1 signifies identical vectors, 0 implies no similarity, and -1 denotes entirely dissimilar vectors. The calculation involves dividing the dot product of the vectors by the product of their magnitudes or norms, ensuring that the similarity measure remains invariant to the lengths of the vectors, depending solely on their directions.

The utility of cosine similarity extends across various domains, including natural language processing, information retrieval, and data mining. It provides a method for quantifying the similarity between vectors or documents based on their corresponding orientations within a multi-dimensional space.

Two bar charts analyze the distribution of scores obtained during the tests at two link prediction thresholds, Tr = 0.9 and 0.95, and two values of neighbors in ego-graphs, L_2 = 5 and L_2 = 2000. Opting for a second choice eliminates de facto limitations on the number of neighboring selections.

The data range is partitioned into four equal segments outlined by data quartiles, each assigned a distinct color for visual clarity: red, yellow, blue, and green. This color scheme highlights specific regions of interest, particularly the red zone at the bottom, which is expected to contain a higher proportion of low-confidence scores, and the green zone at the top, where high-confidence scores are anticipated.

The horizontal axis of each histogram represents the count of instances where a specific number of papers (shown on the vertical axis) were successfully recovered based on the chosen threshold. This allows us to analyze the distribution of recovered citations at different confidence levels and identify potential patterns within the data. The "red" category at the bottom of the graph is predicted to contain more suspected citations, while the "green" category at the top is expected to include consistently cited papers.

The charts portray the count of successfully recovered instances on the horizontal axis, with the associated unnormalized frequencies displayed on the vertical axis. These frequencies signify the number of notes successfully retrieved corresponding to each recovery count. Notably, the categories marked by colors, such as the lowest "red" category (anticipated to contain the most suspected citations) and the highest "green" category (indicative of the most consistent ones), are of primary interest.

4.1.1. Case of L_2 = 5

The corresponding frequencies of the reconstruction edge numbers are given in Table 1.

Table 1. The recovering edge distributions for the CORA dataset for $L_2 = 5$.

Values/*Tr*	0.90	0.95
0	785	1913
1	5993	6293
2	4044	3703
3	2463	2036
4	1370	1060
5	643	458
6	289	182
7	91	50
8	31	19
9	9	4
10	1	1

The quartile values are presented in the subsequent table (Table 2).

Table 2. The quartile values of the recovering edge distribution for the CORA dataset for $L_2 = 5$.

Q/*Tr*	0.90	0.95
*Q*1	0.52	0.32
*Q*2	1.27	0.94
*Q*3	2.39	1.97

Upon comparison of the results presented in Figure 2 and Tables 1 and 2 with those detailed in [10], it becomes clear that the higher recovery threshold (0.95) contributes to the observation and that both distributions demonstrate statistically significant positive skewness. The distribution is skewed towards lower values, characterized by a long right tail with the mean lying to the right of the median.

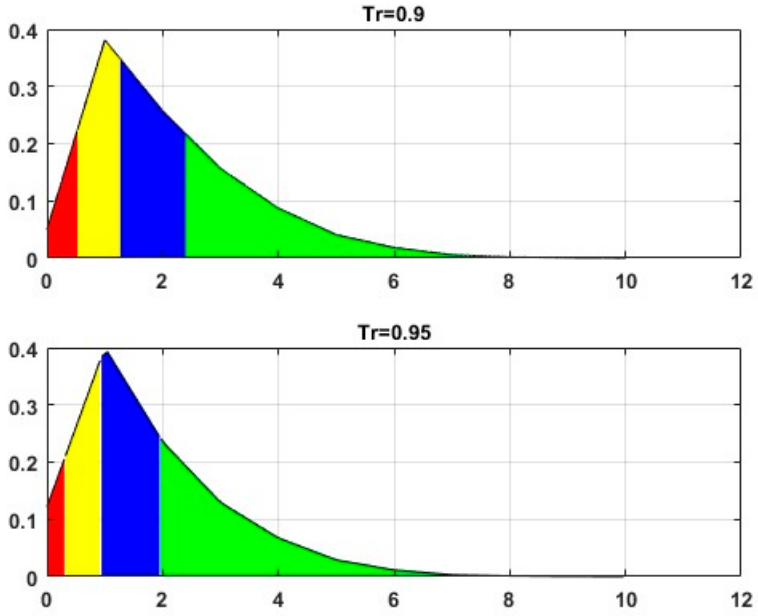

Figure 2. Distributions of edge recovering for the CORA dataset for $L_2 = 5$.

4.1.2. Case of $L_2 = 2000$

As previously said, choosing $L_2 = 2000$ discards the limitation on the number of nearest neighbors.

The associated frequencies are provided in Tables 3 and 4.

Table 3. The recovering edge distributions for the CORA dataset for $L_2 = 2000$.

Values/*Tr*	0.90	0.95
0	170	822
1	6803	5334
2	4551	4060
3	3255	2987
4	2148	1950
5	1184	991
6	476	392
7	191	154
8	49	38
9	8	5
10	2	2

Table 4. The quartile values of the recovering edge distribution for the CORA dataset for $L_2 = 2000$.

Q/*Tr*	0.90	0.95
*Q*1	0.67	0.63
*Q*2	1.54	1.54
*Q*3	2.80	2.78

Upon examination of the bar chart in Figures 2 and 3 and Tables 1 and 2 generated for different L_2 values, a distinct similarity among them becomes evident. The distributions are not significantly different. This observation suggests the presence of a consistent underlying structure within the dataset that remains resilient to variations. Most data point the cluster toward the left side, with a right tail extending further. The overall trend remains unaltered, and this observation highlights the existence of a robust and unwavering underlying structure within the dataset that remains impervious to permutations.

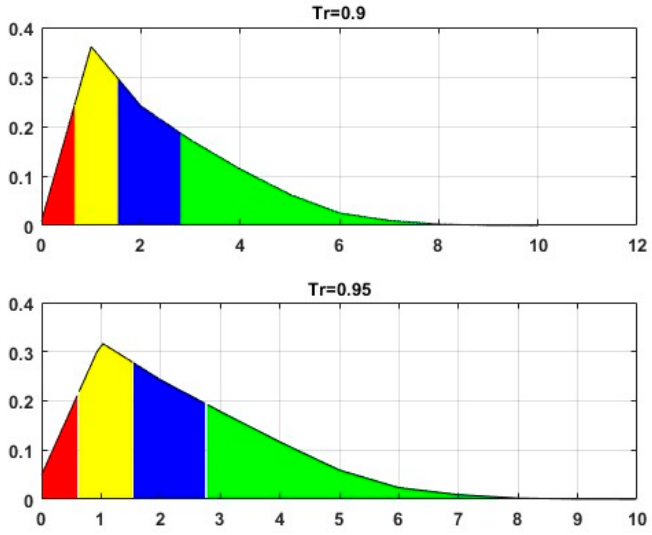

Figure 3. Distributions of edge recovering for the CORA dataset for $L_2 = 2000$.

Given these observations, it becomes interesting to contemplate an experiment designed to uncover the robustness of the approach in identifying spurious citations and to gauge the system's response to adding artificial noise references.

4.1.3. Case of $L_2 = 5$ for a Dataset Noised CORA Version

As a precautionary measure, we opt to examine a perturbed version of the CORA dataset for sanity checks. This involves introducing artificial randomness by adding 20% more edges to the dataset. The following Figure 4 exhibits histograms of the recovered edges.

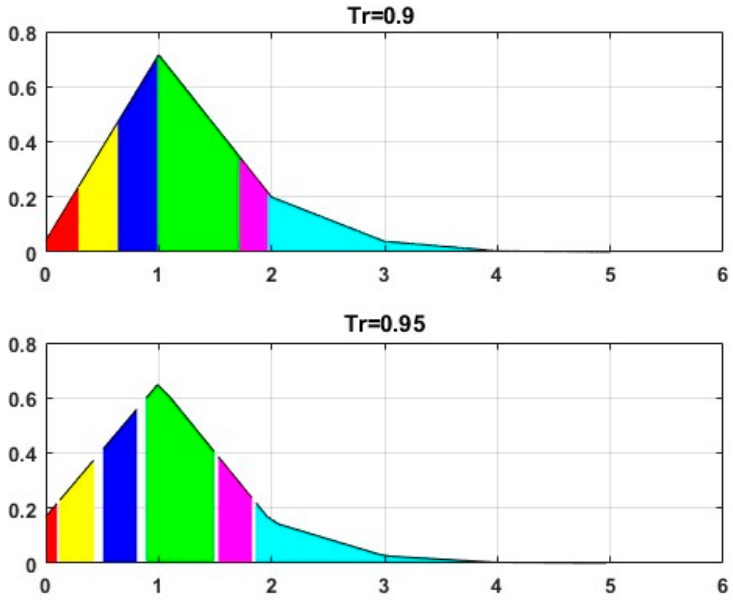

Figure 4. Distributions of edge recovering for the CORA disturbed dataset for $L_2 = 5$.

The colors in this visualization correspond to ten successive deciles, dividing the data into ten equal frequency groups. The associated distributions and percentiles are given in the following Tables 5 and 6.

Table 5. The recovering edge distributions for the CORA disturbed dataset $L_2 = 5$.

Values/*Tr*	0.9	0.95
0	804	3252
1	13,808	12,590
2	3844	2883
3	749	506
4	76	55
5	9	4

Table 6. The percentile values of the recovering edge distribution for the CORA disturbed dataset for $L_2 = 5$.

Percental/*Tr*	0.9	0.95	Colar
*Q*1	0.29	0.12	red
*Q*2	0.64	0.51	yellow
*Q*3	0.99	0.89	blue
*P*90	1.72	1.53	green
*P*95	1.97	1.86	magenta
*P*100	5.00	5.00	cyan

Compared to the preceding experiments, the most recent evaluation reveals a significant decline (approximately 50%) in the central tendency of the variable depicting the quantities of the reconstructed edges. We noted a significant increase in the occurrence of edges reconstructed only once, which closely matched the number of introduced artificial citations. The size of the group reveals that it encompasses not only the added edges but also a significant number of existing true edges whose restoration is compromised by the current network noise. Furthermore, the overall value range of the variable exhibited a predictable decrease, aligning with our expectations.

These observations are reinforced by a distinct and notable downward trend, indicating a stronger concentration of values towards the lower spectrum of the variable's range.

Such a substantial shift strongly implies that the introduction of noise likely impeded the network's capacity to accurately encode and reconstruct edges, leading to a considerable reduction in the quantities of reconstructed edges. Consequently, the proposed methodology effectively captures the distortion within the underlying network structure.

4.2. CiteSeer Dataset

The CiteSeer dataset (e.g., URL: https://paperswithcode.com/dataset/citeseer (accessed on 6 March 2024)) is a well-known and frequently employed academic dataset within information retrieval and machine learning. It is a valuable resource for tasks such as citation network analysis and document clustering, mainly focusing on scientific papers in computer science and related domains. The CiteSeer dataset includes 3312 scientific publications classified into six distinct classes, with a citation network featuring 4732 links. Each publication is represented by a binary word vector, using 0 and 1 to signify the absence or presence of the corresponding word from the dataset's dictionary. The dictionary encompasses a total of 3703 unique words. A partial visualization of this dataset is presented in the following (Figure 5).

Figure 5. Partial visualization of the CiteSeer dataset.

The CiteSeer dataset significantly contributed to advancements in research areas such as information retrieval, machine learning, and natural language processing. Its content and structure provide a foundation for developing and evaluating algorithms for analyzing academic documents. Key attributes of the CiteSeer dataset encompass the following:

- Citation Network: The dataset provides comprehensive information on citation links among different papers, facilitating the examination of citation patterns and relationships between scientific publications.
- Document Metadata: Each document within the dataset is accompanied by metadata such as title, authors, and abstract, supplying essential details for text-based analyses.
- Bag-of-Words Representation: Documents are represented using a bag-of-words model, where the presence or absence of specific words serves as features.
- Clustering and Classification Tasks: Researchers commonly leverage the CiteSeer dataset for tasks like document clustering and classification, aiming to group similar documents or predict document categories.

In comparison with the previous data, the following parameters are changed.

- *N_encoder_layers* = 8 (Number of encoder layers).
- L_2 = 5 (Numbers of neighbors in ego graphs).

The number of encoder layers is determined based on the experiments conducted in [13]. The outcomes of the experiments are presented in Figure 6 and Tables 7 and 8.

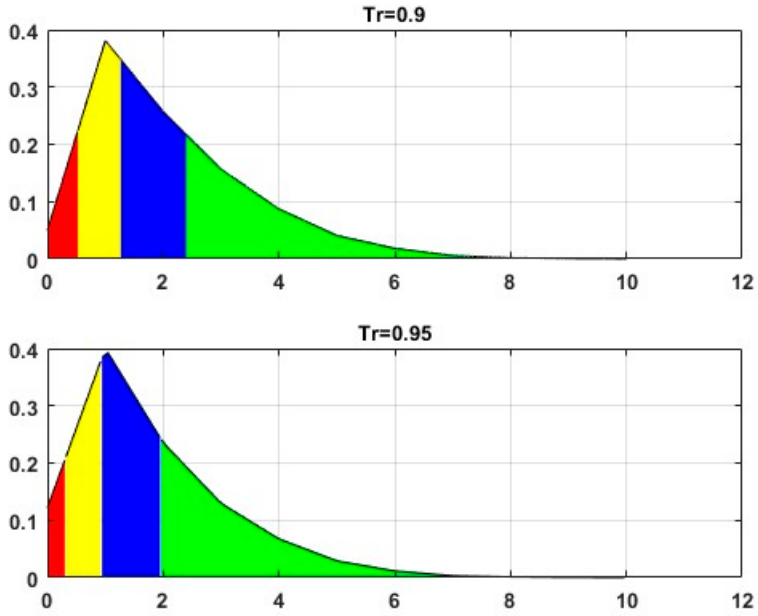

Figure 6. Distributions of edge recovering for the CiteSeer dataset for $L_2 = 5$.

Table 7. The recovering edge distributions for the CiteSeer dataset for $L_2 = 5$.

Values/*Tr*	0.90	0.95
0	19	133
1	8995	8935
2	3054	3019
3	549	533
4	90	87
5	15	15

Table 8. The quartile values of the recovering edge distribution for the CiteSeer dataset for $L_2 = 5$.

Q/*Tr*	0.9	0.95
*Q*1	0.35	0.34
*Q*2	0.71	0.70
*Q*3	1.17	1.16

The results obtained from the CiteSeer dataset exhibit an inherent resemblance in its internal structure to one of the Cora datasets. Notably, the range of reconstructed edges is a bit broader, which could be connected to the CiteSeer dataset's denser configuration.

5. Conclusions

This paper discusses a new attitude to identifying illegitimate citations. The method is built upon the Generalization of Transformer Networks for Graphs and incorporates a masking mechanism to disrupt patterns in altered citation embeddings. Testing validates the approach's efficacy, with masking embeddings provided by the transformer method shining as a dependable tool for uncovering citation manipulation.

While detecting anomalies under regular citation patterns, the model has limitations with multi-disciplinary works and "sleeping beauties"—articles that went unnoticed initially but later experienced a surge in recognition. This phenomenon can arise

from breakthrough discoveries or, simply, later appreciation, challenging this method's detection capabilities.

Around 75% of the total edges (citations) prove susceptible to the distortion procedure, failing to withstand it. The instability of these edges, with their heightened sensitivity to data modifications, sets them apart from the system core's reliable internal structure. Consequently, the associated citations may be deemed dubious and potentially manipulated. This underscores a nuanced dimension of the dataset's integrity and emphasizes the potential impact of specific edges on its structural stability.

The analysis, even though it explores datasets with distinct internal structures, yields sufficiently similar results. This unexpected finding points towards a possible universal inclination within the mutual citation system, suggesting the presence of shared characteristics that transcend the specificities of individual datasets. An intriguing observation is the consistent revelation of a stable core within the citation network across both datasets. While the precise mechanism behind this core's formation remains unclear, it could be linked to the gradual accumulation of reliable links over time. Interestingly, even in datasets like these, which receive regular updates to incorporate newly published articles, as indicated in [10], most edges showcase instability and a lack of relevance. This consistency across different datasets points toward a generalizable property regarding edge reliability, implying that a considerable portion of connections within citation datasets might be less trustworthy or more susceptible to manipulation.

Expanding on this observation, it is imperative to recognize that the positive skewness in the distribution of reconstruction scores signifies a prevailing tendency for data points to lean toward lower scores. With its pronounced right-skewed tail, this unimodal distribution indicates that a substantial portion of the data is concentrated on the left side. At the same time, the mean is disproportionately influenced towards higher scores. Consequently, the prevailing pattern suggests that many references exhibit relatively modest reconstruction scores, prompting consideration of their potential suspicion or manipulation.

The consumption of Graph-Masked Autoencoders (GMAEs) results in a more refined model of citation distribution, capturing the intrinsic connections between papers using additional textual information. This approach distinguishes itself from [10] by uncovering characteristic right-skewed unimodal empirical distributions, indicating a closer alignment with actual citation behavior.

One of the experiments delves into the model's readiness when confronted with an artificially disturbed citation graph, aiming to gauge the approach's trustworthiness. Essentially serving as a sanity check, this assessment validates the model's adeptness in flagging artificially introduced links as highly suspect. The attained findings underscore the model's proficiency in detecting anomalies, affirming its effectiveness and reliability.

The suggested method leverages a stable knowledge core within a graph to track the latest research developments in a specific field. Regularly updating the dataset with new articles and integrating them into the existing knowledge networks provides a dynamic overview of research trends and advancements. Link evaluation for a specific article can be achieved by applying the aforementioned procedure, followed by analyzing the links' position within the general recovery histogram.

The authors are grateful to the anonymous reviewers for their insightful comments, which significantly assisted in improving the quality of this article.

Author Contributions: A collaborative effort brought this project to life. R.A., E.R. and Z.V. collaborated to develop the model and craft the paper. M.B.H. and A.M. designed and executed the experiments. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Prabha, C.G. Some aspects of citation behavior: A pilot study in business administration. *J. Am. Soc. Inf. Sci.* **1983**, *34*, 202–206. [CrossRef]
2. Resnik, D.B.; Gutierrez-Ford, C.; Peddada, S. Perceptions of Ethical Problems with Scientific Journal Peer Review: An Exploratory Study. *Sci. Eng. Ethics* **2008**, *14*, 305–310. [CrossRef] [PubMed]
3. Wilhite, A.; Fong, E. Coercive citation in academic publishing. *Science* **2012**, *335*, 542–543. [CrossRef] [PubMed]
4. Wren, J.D.; Georgescu, C. Detecting anomalous referencing patterns in PubMed papers suggestive of author-centric reference list manipulation. *Scientometrics* **2022**, *127*, 5753–5771. [CrossRef]
5. Dong, M.; Zheng, B.; Quoc Viet Hung, N.; Su, H.; Li, G. Multiple rumor source detection with graph convolutional networks. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management, Beijing, China, 3–7 November 2019; pp. 569–578.
6. Lu, Y.J.; Li, C.T. Graph-aware co-attention networks for explainable fake news detection on social media. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Virtual Conference, 5–10 July 2020; pp. 505–514.
7. Bian, T.; Xiao, X.; Xu, T.; Zhao, P.; Huang, W.; Rong, Y.; Huang, J. Rumor detection on social media with bi-directional graph convolutional networks. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 549–556.
8. Yu, S.; Xia, F.; Sun, Y.; Tang, T.; Yan, X.; Lee, I. Detecting outlier patterns with query-based artificially generated searching conditions. *IEEE Trans. Comput. Soc. Syst.* **2020**, *8*, 134–147. [CrossRef]
9. Liu, J.; Xia, F.; Feng, X.; Ren, J.; Liu, H. Deep Graph Learning for Anomalous Citation Detection. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, *33*, 2543–2557. [CrossRef]
10. Avros, R.; Keshet, S.; Kitai, D.T.; Vexler, E.; Volkovich, Z. Detecting Pseudo-Manipulated Citations in Scientific Literature through Perturbations of the Citation Graph. *Mathematics* **2023**, *11*, 3820. [CrossRef]
11. Grover, A.; Leskovec, J. Node2vec: Scalable Feature Learning for Networks. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, San Francisco, CA, USA, 13–17 August 2016; ACM: New York, NY, USA; pp. 855–864.
12. Zhang, S.; Chen, H.; Yang, H.; Sun, X.; Yu, P.S.; Xu, G. Graph Masked Autoencoders with Transformers. *arXiv* **2022**, arXiv:2202.08391.
13. Dwivedi, V.P.; Bresson, X. A Generalization of Transformer Networks to Graphs. *arXiv* **2020**, arXiv:2012.09699.
14. Ying, C.; Cai, T.; Luo, S.; Zheng, S.; Ke, G.; He, D.; Shen, Y.; Liu, T.-Y. Do transformers really perform bad for graph representation? *arXiv* **2021**, arXiv:2106.05234.
15. Hamilton, W.; Ying, Z.; Leskovec, J. Inductive representation learning on large graphs. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 1024–1034.