Anomalous Citation Detection in Journal Networks: A Comparative Study of CIDRE and GLADC

Arlei Silva
Department of Computer Science
Rice University
2082 Duncan Hall, 6100 Main St, Houston, TX

Jae Jun Ku
Rice University
6100 Main St, Houston, TX
jk75@rice.edu

Cameron Liu
Rice University
6100 Main St, Houston, TX
hl101@rice.edu

Bowen Yao
Rice University
6100 Main St, Houston, TX
by18@rice.edu

Abstract

This paper explores advanced machine-learning technique applied to citation networks to detect anomalous citation behaviors. Utilizing a dataset comprising of a snapshot of the Microsoft Academic Graph (MAG) from 2013, we implement two algorithms, CIDRE and GLADC, to address the challenges posed by citation manipulation. CIDRE focuses on identifying anomalous groups by computing donor and recipient scores, while GLADC leverages contrastive learning to discern irregularities in graph-level structures. Our evaluation highlights the effectiveness of these models in detecting anomalous activities within academic citations, with CIDRE achieving a recall of 0.68 and GLADC achieving a recall of 0.74. Our results demonstrate GLADC is better in detecting citation cartels and has greater potential to enhance the integrity of bibliometric indicators.

1 Introduction

The exponential growth of research publications demands swift and equitable methods for evaluating research outputs. The academic community often relies on citation-based bibliometric indicators[1] like the h-index and the citation numbers to assess the performance of individual researchers, institutions, and national research outputs.[2] Despite their widespread adoption as measures of quality and prestige, these indicators are not without controversy. Increasingly, researchers face pressure to achieve high citation counts which significantly influences a journal's readership and subscription rates. This pressure can lead to beneficial practices aimed at improving publication quality. However, it also fosters malicious practices, including the manipulation of citation numbers through self-citations or the formation of "citation cartels," [3] where groups of journals or researchers engage in a coordinated effort to cite each other's work excessively. [4]

For instance, a notable case in 2011 involved two papers that significantly boosted a journal's impact factor by providing numerous citations to each other.[5] Since then, similar instances have surfaced annually, demonstrating that citation cartels are both easy to establish and difficult to detect.

Addressing these issues necessitates robust methods to detect and address anomalous citation behaviors, thus ensuring the integrity and trustworthiness of academic discourse. However, traditional methods for detecting these cartels often involve identifying densely interconnected nodes within citation networks, which can mistakenly flag normal community citation behaviors as anomalous due to the natural tendency of journals within the same field to cite each other frequently.[6]

This project proposes the application of graph machine-learning techniques to a citation network, framing it as a graph classification challenge, aimed at identifying irregular citation patterns. We employ two cutting-edge algorithms: CIDRE and GLADC.

Anomalous citation groups are inherently difficult to detect with community detection algorithms. CIDRE improves the common clustering algorithm by computing and updating the donor and recipient scores for each journal and removing journals that are neither a donor nor a recipient until no journal is further removed. CIDRE then partitions journals into disjoint groups Ul (l = 1, 2, . . .), where each Ul is the maximal weakly connected component in the network consisting of nodes belonging to U and the residual edges.[7]

Graph-level Anomaly Detection with Contrastive Learning (GLADC) enhances graph-level anomaly detection through contrastive learning, a method that has shown promising results in molecular graph anomaly datasets.[8] As opposed to traditional anomaly detection methods that rely on node-level feature detection (whereby researchers classify nodes within a single graph that are different from the norm), GLADC considers the entire academic graph as a set of smaller communities and considers differences in both local and global graph structures to determine if a journal "group" is anomalous.

By applying these methods, we aim to uncover abnormal structural and feature properties within citation networks. Our evaluation will focus on comparing the effectiveness of these models against established benchmarks to determine their capability to detect "citation cartel" activities more efficiently than existing approaches.

2 Methodology

2.1 Dataset

The main dataset we use in this project is a snapshot of the Microsoft Academic Graph (MAG) from 2013. Edges are present between journals (nodes) that cite other journals, and edge weights represent the number of citations made to the papers in the prior two years.[9] The dataset encompasses bibliographic information from 8,038,733 papers across 31,385 journals in various research fields. We have the node file[1] and the edge file[2]. This information includes journal names, publication years, references, author details, and what community the journal belongs to.

Each journal is represented as a node in a network, where edges between nodes signify citations between journals. The weight w_{ij} of an edge from journal i to journal j in year t is defined by the number of citations from papers in journal i to papers in journal j, specifically within the two years before t ($t - 2$ to $t - 1$), aligning with the time window used for calculating the Journal Impact Factor.

Additionally, when preparing data for the GLADC experiment, a database of journals that were suspended by JCR (data provided from Clarivate Analytics)[3] was used to determine anomalous graphs within the set of journal graphs. An anomalous graph is classified as any graph containing at least one of the suppressed journals from 2013.

Table 1 describes the properties of the complete MAG, as well as three other datasets (BZR, DHFR, COX2) that were examined in the GLADC paper. MAG is the dataset representing the 2013 snapshot of the Microsoft Academic Graph. BZR, DHFR, and COX2 are molecular graph datasets. Compared to the other datasets, MAG exhibits a substantially higher number of graphs (representing academic communities), a lower number of average nodes (representing journals) per community, and a substantially higher number of average edges (representing citations) per community. Most significantly, the anomaly ratio- representing the ratio between the number of anomalous graphs and the number of total graphs in the dataset- is extremely small for MAG compared to the other datasets.

2.2 CIDRE

The CIDRE algorithm is designed to detect anomalous groups of journals in citation networks, categorizing journals as either donors or recipients based on their citation behaviors. It quantifies the extent to which a journal functions as a donor or recipient within a group by using donor and recipient

[1] Node File: https://raw.githubusercontent.com/skojaku/cidre/main/data/journal-citation/community-label.csv

[2] Edge File: https://raw.githubusercontent.com/skojaku/cidre/main/data/journal-citation/edge-table-2013.csv

[3] Suspended Journals: https://tinyurl.com/mry9zn3r

Table 1: Statistics and Properties of Graph Anomaly Datasets

Datasets	MAG	BZR	DHFR	COX2
Avg. Nodes	22.08	35.75	42.43	41.22
Avg. Edges	1136.92	38.36	44.54	43.45
Graphs	1918	405	467	467
Anomaly ratio	1.61%	21.23%	39.03%	21.84%

scores. These scores are calculated based on the journal's citation out-strength and in-strength relative to expected values from a null model, adjusting for excessive citation behaviors identified by a specified threshold.

First, the model assesses whether citations are excessive by examining the proportion of recent citations and comparing actual citation counts to those predicted by a null model. If citations are deemed excessive, a p-value is calculated to statistically validate this observation. The algorithm uses a Benjamini-Hochberg correction to control for false discoveries across multiple comparisons, ensuring that only the most statistically significant edges (citations) are considered.

To isolate groups of journals with anomalous citation patterns, CIDRE employs a method similar to the k-core decomposition algorithm. This involves pruning the citation network to retain only significant edges, initializing potential anomalous groups, and iteratively refining these groups by recalculating the donor and recipient scores and removing journals that do not meet the score thresholds. The final groups are defined as maximal weakly connected components in the pruned network, with additional filtering based on the total weight of within-group citations to ensure the groups are sufficiently dense.[7]

2.3 GLADC

The GLADC (Graph Level Anomaly Detection with Contrastive Learning) framework is a recently developed deep learning model designed to identify anomalous graphs within a set by examining both local and global aspects of graph anomalies. This model uses a dual-graph encoder module to encode and decode graphs for generating node-level representations while enhancing graph-level representations through contrastive learning, without using traditional data augmentation techniques like node dropping or edge perturbation. This avoids introducing noise into the graph data and preserves the structural properties of the graphs.[8]

The framework first uses a graph convolutional autoencoder to learn node-level representations, which are then used to calculate the reconstruction loss combining both structural and attribute reconstructions of the graph. GLADC then employs a dual encoder that perturbs one of the graph encodings with Gaussian noise, followed by a graph max pooling to summarize the node features into graph-level features. This process allows the model to contrast graph representations using a specially designed loss function, which aids in distinguishing between normal and abnormal graphs by comparing the error in representations of the real and reconstructed graphs.

GLADC is first trained on only normal graph data to establish baseline representations and then tested on a mix of normal and abnormal graphs to assess the anomaly detection capability. The model identifies anomalies by measuring the discrepancy in node-level and graph-level representations between the input and its reconstructed counterpart, quantifying this discrepancy as an anomaly score.

3 Main Challenges

3.1 CIDRE

Non-machine learning approaches like the Citation Donors and Recipients (CIDRE) algorithm [7] demonstrate reasonable effectiveness. However, CIDRE relies on statistical models that may not fully capture the intricate dynamics of citation networks. The development of CIDRE utilized citation network data from the now-discontinued Microsoft Academic Graph. Additionally, in our reproduction efforts of the CIDRE algorithm, we encountered deprecated functions within its Python implementation, and we needed to manually update the package functions.

CIDRE operates by first eliminating non-active nodes (neither donors nor recipients) from the graph, then segmenting the graph into maximal weakly connected components, and examining each group for $\theta > 50$. This method assumes that anomalous citations occur within their respective clusters, reflecting the commonality of citations among papers in similar research fields. Nevertheless, this could be circumvented by malicious entities aware of the algorithm's mechanics, potentially orchestrating undetectable collusion across disparate fields. Also, the value of θ is a hyperparameter we need to fine-tune. But since each group has a different number of weighted edges and nodes, ideally, θ value should be different for each anomalous group.

3.2 GLADC

3.2.1 Data Preparation

In the process of collecting data to train the GLADC model, we discovered that the Microsoft Academic Graph which was used in the original CIDRE paper was discontinued in January 2022. We were able to acquire a snapshot of the Microsoft Academic Graph from 2013 from the authors of the CIDRE paper that was used in our experiments; however, ideally, a more recent dataset would have yielded more relevant results.

In implementing the GLADC algorithm, we faced a challenge with the dataset lacking node attributes or labels, a requirement not met in the example datasets utilized by the GLADC code. To overcome this, we generated node features based on node degree information, following the recommendations for plain graphs described in the literature. Each node was labeled with the sum of edge weights, representing the total citation count it received. This approach provided a simple way for the GLADC algorithm to gain information about local structures within different graphs.

3.2.2 Computational Power

Compared to the molecular datasets used in the GLADC paper, MAG exhibited a significantly higher number of edges per graph. When loading the data into a standard cloud computing provider for deep learning such as Amazon EC2 and Google Colaboratory, we discovered that MAG needed around 25x more memory for storage compared to datasets such as BZR, COX2, and DHFR. This far exceeded the maximum system RAM that we had access to (50GB).

As such, we created scripts to prune the original MAG data by around 70% during data processing, which was done by random sampling of graphs within the dataset. The new dataset was then re-evaluated to ensure that it maintained a representative balance of the original characteristics, such as the distribution of node degrees and edges, and anomaly ratio.

After pruning, we noticed that the model was unable to detect anomalies effectively during the testing phase, as nearly all of the anomalous graphs had been filtered out during random selection. In order to solve this issue we artificially introduced anomalous graphs into the test set and used undersampling to decrease the number of the non-anomalous graphs in the test set. This also brought the anomaly ratio of the pruned MAG dataset to 6.31%, which was closer to the ratios of the other molecular datasets that GLADC had performed well on.

3.2.3 Deprecated Libraries and Dependencies

We also encountered significant challenges due to deprecated libraries and dependencies within the starting code. Although the GLADC paper was published only 2 years ago, many of the Python libraries present in the requirements.txt file were using older versions and relied on other deprecated libraries. For example, the Torch library had major updates to several functions that were not compatible with the code. As such, we manually reviewed all of the dependencies, and in cases where directly updating the package led to unsolvable errors, we employed other methods or libraries to achieve the same task.

4 Evaluation

4.1 Experimental Implementation

4.1.1 CIDRE

In our experiments, we utilized the implementation provided by the CIDRE paper. The threshold parameter θ in the CIDRE model defines the minimum proportion of excessive citations that journals either give as donors or receive as recipients within their groups. The groups that persist at higher θ values represent more closely knit citation networks, indicating stronger candidates for citation cartels. For this study, we set θ = 0.01 to 0.15 for the experiment. After the detection by CIDRE, We regard each group with more than $\theta_w = 50$ within-component citations as an anomalous citation group ($\theta_w$is the sum of the weights of the non-self-loop edges within the group).

4.1.2 GLADC

We used a modified implementation of the GLADC code provided by the authors of the GLADC paper. PyTorch was used for creating the neural network, which consisted of two layers of Graph Convolutional Networks (GCNs). Each layer in the graph encoder module, shared by both the dual-graph encoder module and the standard graph encoder, had dimensions set sequentially, with a hidden layer dimension of 256 and an output layer dimension of 218. We used the Adam optimizer during training with a learning rate of 0.0001. The batch size was set at 300, and the number of epochs at 100. The loss was calculated using a function that combined reconstruction loss from the autoencoder and contrastive loss. The contrastive loss was implemented using a temperature-scaled cross-entropy loss function. This function uses a temperature parameter, set at 0.2, which adjusts the sharpness of the distribution of the softmax output, controlling the separation between positive and negative pairs in the learned embedding space.[8] Finally, the GLADC model was run on the condensed MAG dataset with 5-fold k cross-validation.

4.2 Experimental Results

CIDRE yielded an acceptable result when tested on the dataset. CIDRE does not label each node in the graph with a probability of being anomalous; instead, it iteratively removes nodes and edges from the graph and labels the remaining nodes all as anomalous once no journal is further removed. We could argue this is a slightly less ideal design than GLADC because true anomalous nodes may be removed during this process. As shown in Figure 3, we experimented with different theta values on the model, and the best performance we got was that we detected 23 of 34 anomalous journals reported by JCR in 2013, which is a true positive rate of 0.676. But we can also see more journals and groups are labeled as anomalous as theta decreases, which can lead to more human labor for inspection.

Compared to CIDRE, GLADC yielded promising results. To evaluate the performance of the GLADC model on the dataset, we used AUC-ROC (Area Under the Receiver Operating Characteristic Curve), a common method for evaluating performance on anomaly datasets [8]. Figure 1 depicts the Receiver Operative Characteristic (ROC) curve for the model, averaged over all 5 k-fold validation sets. In our experiment, the GLADC model achieved an AUC-ROC score of 0.83, reflecting a high level of accuracy in anomaly detection within the citation network. This score demonstrates the model's effectiveness, especially considering the MAG dataset's complexity and low anomaly ratio.

We also wanted to explore the effects of changing the graph representation dimensions on model effectiveness. As seen in Figure 2, the model achieves the best performance with a representation dimension size of 128 and 512. This matches the findings of the original GLADC paper, where different molecular datasets also achieved the best performance when the dimension size was set at 128, and the worst at 256.[8]

4.3 Model Comparison

In comparing the two models' performance in Table 2, we first notice that the dataset is extremely unbalanced, comprising of 31,385 nodes, 34 true labels, and 31,351 false labels. Therefore, metrics like accuracy would be misleading, as both methods would exhibit very high values due to the high proportion of false labels. Instead, we compare the true positive rate (recall) of the two models and

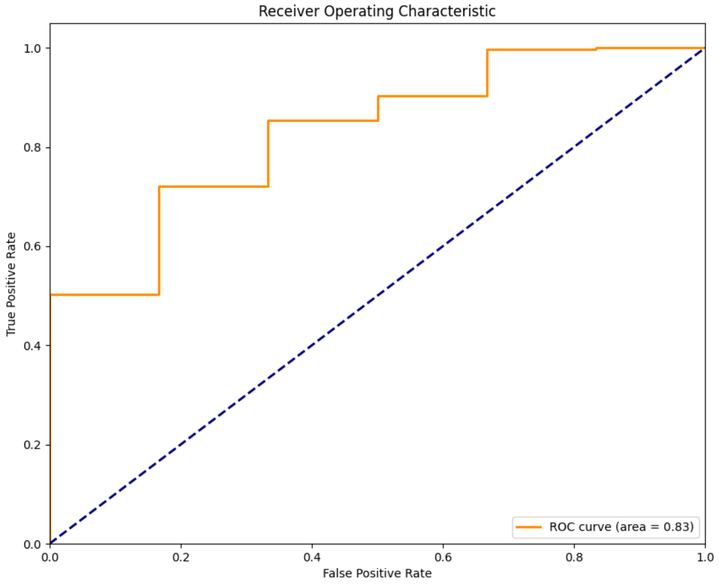

Figure 1: Receiver Operating Characteristic (ROC) curve for the GLADC model.

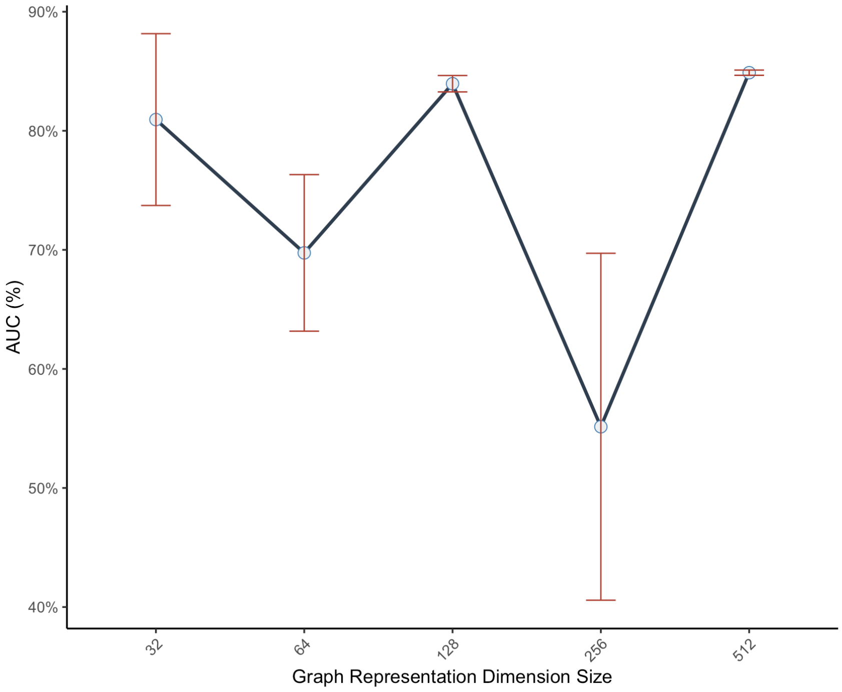

Figure 2: AUC-ROC Over Varying Graph Representation Dimension Sizes

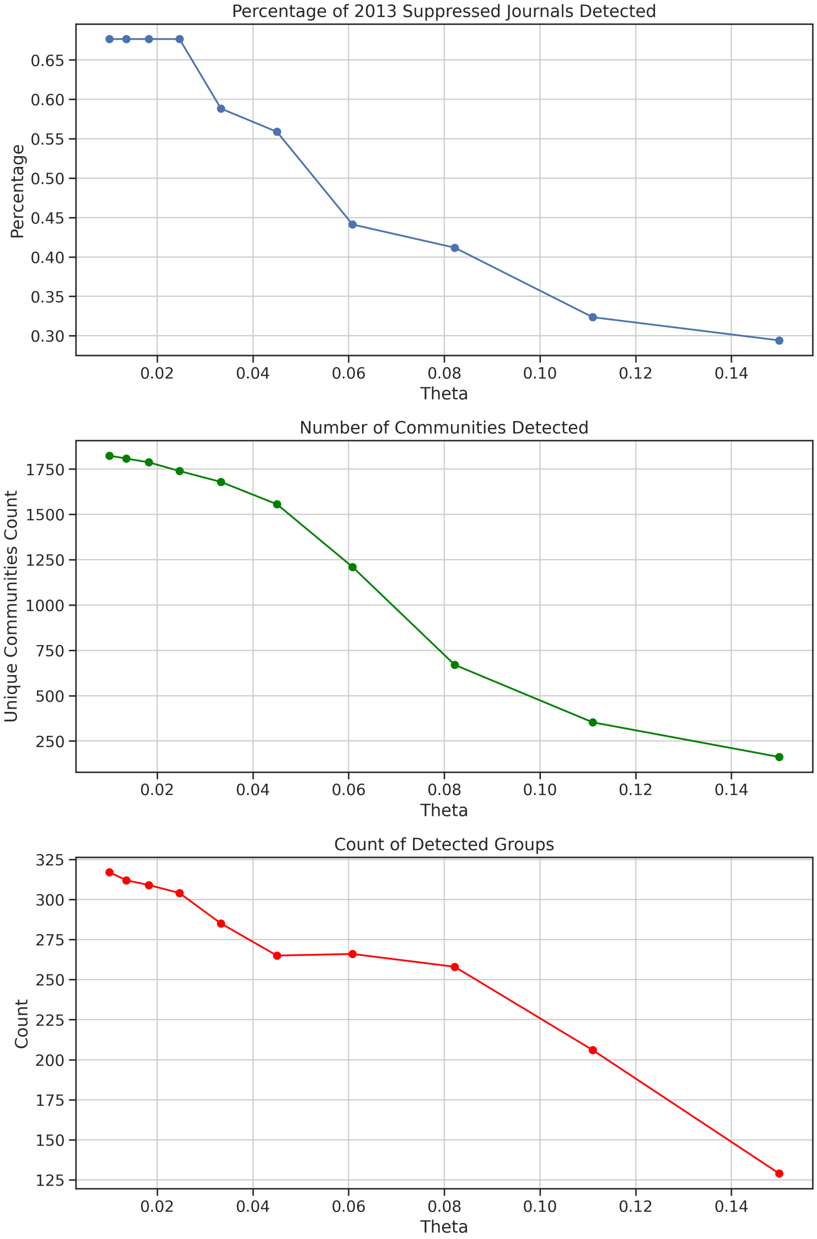

Figure 3: Analysis of Theta Values

Table 2: Comparing CIDRE vs. GLADC

Detection Model	CIDRE	GLADC
True Positive Rate	0.676	0.739
Anomalous Communities Detected	1739	176
True Anomalies (2013)	34	34

the total number of positives detected. As shown in Table 2, GLADC is more effective at identifying anomalous communities than CIDRE. GLADC has a higher true positive rate, as it correctly predicts 73.9% of the suspended journal communities. Additionally, we see GLADC labels significantly fewer communities as anomalous while retaining high TPR, which means GLADC detected a significantly smaller number of false positives than CIDRE, showing that it is more robust against false detection.

5 Further Discussion

Detecting anomalous citations is inherently challenging. Although we can compare our results against journals suspended from JCR, such suspensions may arise from factors other than excessive citations, including publication irregularities, ethical breaches, or manipulation of the peer review process. Moreover, journals suspended by JCR represent only a small fraction of all journals potentially inflating their Journal Impact Factor (JIF) through excessive citations. Consequently, the ground truth may never be revealed.

This makes the task of detecting citation anomalies not only a technical challenge but also a conceptual one, as the very definition of what constitutes an anomalous citation can be subjective and vary across different academic communities. Citation networks are complex, and communities identified by GLADC or CIDRE as "anomalous" may not be a result of malicious intent but other factors such as academic/editorial relationships and collaborative networks.

Still, we find great promise in GLADC as a machine-learning-based method to detect citation cartels. Future iterations of GLADC could benefit from integrating node features that include metadata such as editorial history (the number of reviewers or the revision cycle length), citation intent, and temporal data. Additionally, enhancing the dataset used for training such models is crucial. While the 2013 Microsoft Academic Graph was used for this report, more comprehensive and up-to-date datasets would allow GLADC to train on a broader spectrum of citation activities, including those from less studied or newly emerging fields.

6 Code

The code used to obtain the results for the GLADC model in this project can be found here.

The code used to obtain the results for the CIDRE model in this project can be found here.

References

[1] S. Saha, S. Saint, and D. A. Christakis. Impact factor: A valid measure of journal quality? *Journal of the Medical Library Association*, 91:42, 2003.

[2] Lutz Bornmann and Johann Bauer. Which of the world's institutions employ the most highly cited researchers? an analysis of the data from highlycited.com. *Journal of the Association for Information Science and Technology*, 66:2146–2148, 2015.

[3] G. Franck. Scientific communication—a vanity fair? *Science*, 286:53–55, 1999.

[4] IJ Fister, I Fister, and M Perc. Toward the discovery of citation cartels in citation networks. *Frontiers in Physics*, 4:49, 2016.

[5] P. Davis. The emergence of a citation cartel. https://scholarlykitchen.sspnet.org/2012/04/10/emergence-of-a-citation-cartel/, 2012.

[6] A. Lancichinetti and S. Fortunato. Consensus clustering in complex networks. *Sci. Rep.*, 2:336, 2012. arXiv:1203.6093.

[7] S. Kojaku, G. Livan, and N. Masuda. Detecting anomalous citation groups in journal networks. *arXiv preprint arXiv:2009.09097*, 2020.

[8] X. Luo, J. Wu, J. Yang, S. Xue, H. Peng, C. Zhou, H. Chen, Z. Li, and Q. Z. Sheng. Deep graph level anomaly detection with contrastive learning. *Scientific Reports*, 12(1), 2022.

[9] A. Sinha et al. An overview of microsoft academic service (mas) and applications. In *Proceedings of the 24th International Conference on World Wide Web, WWW'15 Companion*, pages 243–246, New York, NY, USA, 2015. Association for Computing Machinery.