Accumulated Local Effects for Link Prediction with Graph Neural Networks

Paulina Kaczyńska *
University of Warsaw
IPPT PAN

Julian Sienkiewicz
Warsaw University of Technology

Dominik Ślęzak
University of Warsaw

Abstract

We investigate how Accumulated Local Effects (ALE), a model-agnostic explanation method, can be adapted to visualize the influence of node feature values in link prediction tasks using Graph Neural Networks (GNNs), specifically Graph Convolutional Networks and Graph Attention Networks. A key challenge addressed in this work is the complex interactions of nodes during message passing within GNN layers, complicating the direct application of ALE. Since a straightforward solution of modifying only one node at once substantially increases computation time, we propose an approximate method that mitigates this challenge. Our findings reveal that although the approximate method offers computational efficiency, the exact method yields more stable explanations, particularly when smaller data subsets are used. However, the explanations produced with the approximate method are not significantly different from the ones obtained with the exact method. Additionally, we analyze how varying parameters affect the accuracy of ALE estimation for both approaches.

1 Introduction

This study investigates the application of Accumulated Local Effects (ALE) [1], a model-agnostic explanation method, to Graph Neural Networks (GNNs) trained for link prediction. ALE visualizes the impact of a specific feature's value on the model's prediction. Unlike GNNExplainer [2] and PGExplainer [3], which highlight important subgraphs and feature subsets for predictions, or counterfactual methods like GCFExplainer [4], which identify minimal graph alterations to change predictions, ALE provides a different perspective on model behavior. By focusing on individual feature effects, ALE could serve as a valuable complementary tool for GNN explainability, particularly given its broader applicability beyond the GNN domain.

ALE calculation involves modifying specific feature values and assessing the model's predictions on this altered dataset to measure the impact of these changes. While this process is straightforward for tabular data, where multiple points can be modified simultaneously, it presents unique challenges for GNNs. During message passing GNN layers update the node's embedding with the information about its neighbors, which is passed along the edges and later aggregated [5]. Hence, a prediction made for the nodes is influenced by their neighbors. If, during ALE calculation, many nodes would be modified simultaneously, they could influence each other's prediction in an undesirable way. On the other hand, modifying nodes one by one is highly time-consuming.

This work tries to estimate the scope of the aforementioned effect and answer the question: does ignoring this effect and calculating the prediction as it is done with tabular data significantly affect the explanation?

*Correspondence to: pm.kaczynska@student.uw.edu.pl

ATTRIB Workshop at 38th Conference on Neural Information Processing Systems (NeurIPS 2024).

We obtain the ALE estimation both by ignoring and accounting for this effect, and then compare the results from these two approaches. Additionally, we calculate ALE for various parameter sets to analyze how these parameters influence the accuracy of the ALE estimation.

The code used to produce the results described in this work can be found at https: `//github.com/Kaczyniec/ALE-and-GNNs`.

2 Accumulated Local Effects

Accumulated Local Effects (ALE) plots provide a way to visualize the effect of a feature on the predictions of a machine-learning model by accumulating local changes in the predictions as the feature values vary [1]. It is an alternative to Partial Dependence Plots [6] and addresses some of its limitations, such as the sensitivity to feature correlations and the inability to accurately capture interactions between features. The core idea is to measure the local effect of a feature by looking at the changes in predictions when the feature value changes slightly, and then accumulating these changes across the range of the feature. In this expression, the derivative $f^S(X_S, X_C)$ represents the local effect of X_S on the model prediction, and this effect is accumulated over the range from $x_{\min,S}$ to the current value x_S:

$$g_{S,ALE}(x_S) = \int_{x_{\min,S}}^{x_S} \mathbb{E}[f^S(X_S, X_C)|X_S = z_S]dz_S - \text{constant} \tag{1}$$

The empirical estimation of Accumulated Local Effects is given by:

$$\hat{g}_S(x_S) \equiv \sum_h \frac{1}{n_S(h)} \sum_{\{i:x_{i,S}\in N_S(h)\}} [f(z_{h,S}, x_C) - f(z_{h-1,S}, x_C)] \tag{2}$$

In this formula, the summation aggregates the local differences in the model's predictions as the feature X_S transitions from one interval to another. Values $z_{h,S}$ and $z_{h-1,S}$ correspond to the border of the feature x_S interval, and $n_S(h)$ represents the number of observations within the hth interval.

Originally, ALE is centered by subtracting ALE [1] averaged over all possible values of the feature, making it easier to interpret the contribution of each feature relative to its average effect. For the sake of simplicity of analysis, this procedure will not be applied here.

3 Methodology

3.1 Modification of ALE for link prediction

In the task of link prediction, the model returns the probability of an edge existing between two given nodes, v and u. This requires a slight adjustment of the ALE method. Instead of modifying features for both nodes involved in the potential link, we focus on altering the features of only one node, which we designate as v. The other node, u, remains unmodified. In this way, ALE visualizes the effect of the node feature's value on the existence of edges between the modified node and the rest of the dataset.

Graph datasets can be large. Due to this, averaging across all of the nodes present in the dataset could not be feasible. Hence, we take only a subset of size m of nodes, for which we modify the feature X_S we are interested in. We then choose the subset U of size k of nodes, against which we evaluate the link probability for each modified node.

Estimation of ALE from Eq. (2) is modified in Eq. (3) in order to account for the link prediction task and averaging over only a subset of the nodes:

$$\hat{g}_S(x_S) \equiv \sum_h \frac{1}{k} \sum_{u\in U} \frac{1}{m} \sum_{\substack{v(x_{i,S},x_{i,C}):\\ x_{i,S}\in N_S(h)}} [f(v(z_{h,S}, x_{i,C}), u) - f(v(z_{h-1,S}, x_{i,C}), u)] \tag{3}$$

The sum over v is taken over the nodes with X_S in the interval h. The middle sum (which did not appear in Eq. (2)) is taken over nodes u, which can have an edge with v. It is divided by the number of these nodes.

Algorithms To explore the effect of nodes' interaction during message passing on explanation, we implement two versions of ALE. In the first "approximate" version, the node features are treated as the tabular dataset, and for one interval, the model's prediction is computed simultaneously. The fact that they influence each other while message passing is ignored. This version is further called the approximate version. In the second "exact" version, the value of the explained feature is changed for each examined node at a time, in isolation from the other nodes. The algorithms are presented in the Appendix A.

ALE parameters Due to the computational constraints, we calculated the explanations for values of parameters k and m being the power of 2 between 16 and 1024.

'Gold standard' There exists a need for some form of a gold standard to which single explanations could be compared.

The most accurate estimation of ALE was created by averaging the exact explanation for different values of k and m. For every value of the parameters, the intervals (and the first sum in Eq. (3)) remain the same. The latter averages cannot be simply added since the sum of averages is not necessarily the average of sums. However, if the ALE is multiplied by the km, only the sums remain, and the expression becomes additive. In this way, we can sum the predictions obtained during every run of the experiment. We divide it by the number of all predictions in the interval and, in this way, obtain ALE combined from multiple small runs.

In result, the following formula is obtained:

$$g_S(x_S) = \frac{\sum_i k_i m_i g_{S,k_i,m_i}(x_S)}{\sum_i k_i m_i} \tag{4}$$

In this way, multiple ALE profiles can be aggregated into one ALE profile corresponding to the ALE, which would be obtained if predictions from multiple runs were calculated during one ALE estimation.

4 Datasets

We used two datasets. The first one is a citation network of 159,734 Artificial Intelligence research papers from the S2ORC corpus [7] with 227,565 citations between them, enriched with author affiliation data from OpenAlex [8][9]. The second dataset is CD1-E_no2 - a 3D vessel graph of mouse brain vasculature containing 1,664,811 nodes and 2,150,326 edges [10]. In the citation dataset, the explained feature is the fraction of authors affiliated with Big Tech companies, allowing for an investigation into the influence of private sector affiliation on citation patterns. In the vessel graph dataset we explain the z-coordinate of nodes to explore the relationship between brain height and vessel connectivity.

5 Models

The models were trained for the task of link prediction: given two nodes, the model should return if there exists a link between them. A GNN encoder, either two 256-dimensional layers of Graph Convolutional Network (GCN) [11] or Graph Attention Network (GAT) [12], was used to obtain node embeddings, and the dot product of the embeddings was calculated to predict link probability via a sigmoid function. Binary cross-entropy was used as the loss function, with batch normalization [13] applied after both layers. The models were implemented using PyTorch Geometric [14].

The negative sampling of edges was performed, with number of negative samples equal to the number of positive samples. The Citations dataset models were trained for 15 epochs on a CPU, while the CD1-E_no2 dataset was trained for 50 epochs on a GPU.

6 Results

χ^2 Test To determine whether the ALE curves from both methods differ or if they can be used interchangeably, we applied a χ^2 test adjusted for comparing arbitrary curves [15]. The null hypothesis

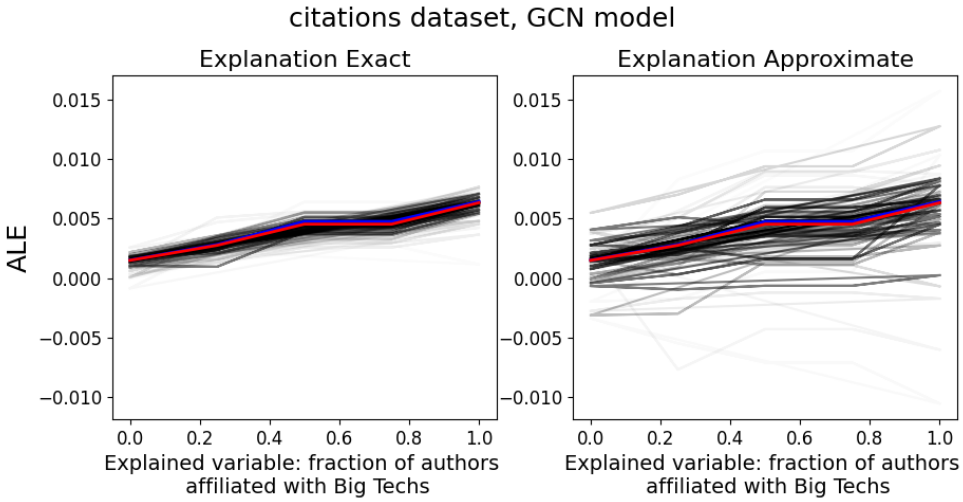

Figure 1: ALE curve calculated for the fraction of authors affiliated with Big Techs. The hue corresponds to the number of edges taken into account during calculating ALE profile. The red line is the 'gold standard' - the average of the exact predictions weighted with the number of predictions taken into account (see Appendix 3.1), and the blue line is the weighted average of approximate predictions.

assumed that the ALE profiles came from the same distribution. As recommended by Hristova and Wimley [15], the degrees of freedom were set to the number of points in the curve. At a significance level of $\alpha = 0.05$, the null hypothesis would be rejected if the χ^2 value exceeded 11.07. For the Citations dataset, the χ^2 values for the ALE curves were 7.165 for the GCN model and 5.413 for the GAT model. For the CD1-E_no2 dataset, the χ^2 values were 17.439 for the GCN model and 1.296 for the GAT model. A statistically significant difference between the curves was observed only for the GCN model trained on the CD1-E_no2 dataset.

Permutation Test We conducted a permutation test to assess whether the exact and approximate explanations differ significantly. The null hypothesis stated that both groups were sampled from the same distribution. The test statistic was the weighted average ALE profile, and the difference was measured as the root mean squared error between the averaged profiles of the two groups. The p-value was the percentage of tests where the difference between the test statistics of the two groups exceeded that of the original group split. A total of $n = 10,000$ splits of ALE curves into two groups were randomly generated.

For the Citations dataset, the p-value for explanations of GCN model was 0.407, and for GAT model - 0.898. For the CD1-E_no2 dataset, it was 0.195 for GCN model and 0.155 for GAT model. No p-value was smaller than the significance level $\alpha = 0.05$. Hence, the null hypothesis stating that samples are taken from the same distribution was not rejected in any case.

Parameters' impact There exists a bigger variability for the single runs of approximate ALE than exact ALE. It can be observed in Fig. 1, 6, and 7. The smaller the number of nodes taken into account, the stronger this effect.

Fig. 2 and 3 show that the higher the k parameter, the smaller the RMSE between approximate ALE and the 'gold standard'. However, this effect is not visible for the exact ALE. From the same plots, we conclude that the higher the time of explanation (proportional to the m parameter), the better the exact ALE.

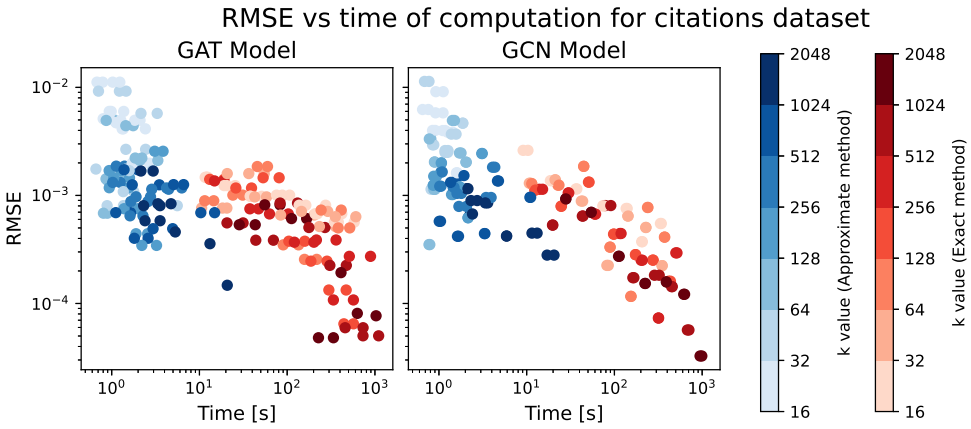

Figure 2: RMSE between runs and 'gold standard' plotted against time of explanation for Citations dataset. The red dots are the exact ALE and the blue dots are the approximate ALE. The hue corresponds to k. The time of exact explanation is roughly proportional to the m parameter.

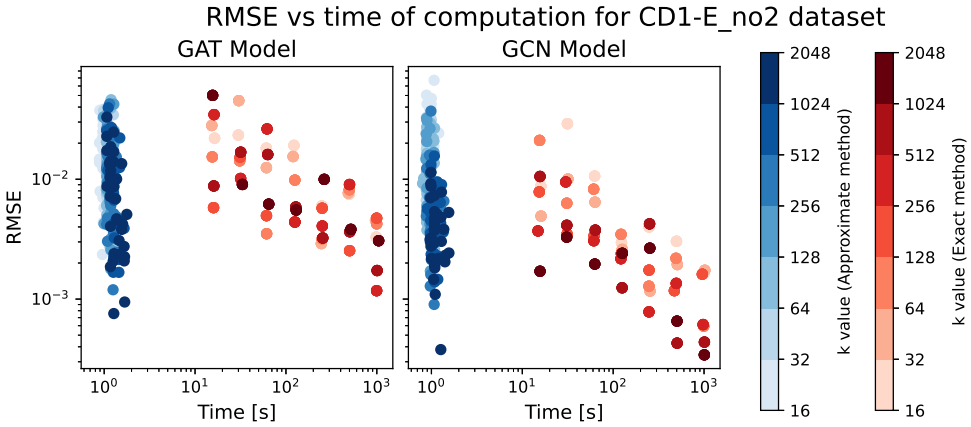

Figure 3: The Root Mean Square Error between runs and 'gold standard' plotted against time of explanation for CD1-E_no2 dataset. The red dots are the exact ALE and the blue dots are the approximate ALE. The hue corresponds to k. The time of exact explanation is roughly proportional to the m parameter. This explanations were performed on GPU.

7 Discussion

The results of χ^2 tests showed that in 3 out of 4 different models, the results obtained with both methods were not significantly different. The permutation test did not show differences between the approximate and exact methods' results in any case.

For the exact method of ALE calculation in link prediction tasks in GNNs, it is more beneficial to increase the m parameter than k parameter. In this way, more nodes with modified feature's value are taken into account. This comes at the time expense since computation scales linearly with the number of nodes in the interval.

Although the approximate method of explanation has greater variability between single runs, it could be used in time-sensitive scenarios. This variability can be reduced by increasing the predicted number of edges (by increasing k and m parameters).

7.1 Relationship between explanations and real-life phenomena

Fig. 1 and 6 show that the probability of being cited increases with the fraction of authors affiliated with Big Techs. This relation between the article's popularity and authors' affiliation is consistent with the literature on this topic [16]. However, it does not support the conclusions of the PageRank and node degree analysis of Giziński et al. [9], which revealed that the most popular were articles with authors affiliated both with Big Techs and Academia. The latter analysis was performed on the Citations dataset, which is also used in this work. This discrepancy between the two analyses could stem from differences in the methodologies used or from the possibility that our models did not capture more complex relationships present in the data.

Additionally, Fig. 7 shows that the probability of a link forming increases with the z-coordinate, whereas Fig. 8 suggests the opposite—a decreasing probability. This contradiction may arise from the two models learning opposing relationships between the z-coordinate and the likelihood of an edge forming between nodes.

These results underscore the utility of ALE (Accumulated Local Effects) in assessing how node features influence GNN predictions. However, it's important to note that explainability tools like ALE reveal only what the models have learned, not necessarily the underlying real-world phenomena.

8 Conclusions

We show how, in most cases, the explanations produced with the approximate method are not significantly different from the explanations produced with the exact method. This leads us to the conclusion that, especially in time-sensitive situations, the node interaction effects can be ignored. However, the exact explanations are usually not only more accurate but also more stable.

The k parameter - the number of nodes possibly having an edge with a modified node - has a clear impact on the accuracy while using the approximate method, but we do not observe a similar impact with the exact method.

8.1 Limitations

The scope of models was restricted to the link prediction task, excluding other tasks such as node classification, edge classification, or graph classification. Additionally, only two GNN architectures were evaluated.

ALE was applied exclusively to continuous variables, limiting its applicability to categorical node features. This restricts the analysis to a smaller number of node features in graph datasets.

The datasets are rather sparse. The interaction between modified nodes should intuitively increase with the density of the graph.

We did not provide formal proof for the observed phenomena. The approximation effects could potentially be influenced by the network's density and the number of message-passing layers, though this relationship was not analytically explored.

References

[1] Daniel W. Apley and Jingyu Zhu. Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(4):1059–1086, September 2020. ISSN 1369-7412, 1467-9868. doi: 10.1111/rssb.12377.

[2] Rex Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. Gnnexplainer: Generating explanations for graph neural networks. 2019. doi: 10.48550/ARXIV.1903.03894. URL https://arxiv.org/abs/1903.03894.

[3] Dongsheng Luo, Wei Cheng, Dongkuan Xu, Wenchao Yu, Bo Zong, Haifeng Chen, and Xiang Zhang. Parameterized explainer for graph neural network. 2020. doi: 10.48550/ARXIV.2011.04573. URL https://arxiv.org/abs/2011.04573.

[4] Mert Kosan, Zexi Huang, Sourav Medya, Sayan Ranu, and Ambuj Singh. Global counterfactual explainer for graph neural networks. 2022. doi: 10.48550/ARXIV.2210.11695. URL https://arxiv.org/abs/2210.11695.

[5] Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. (arXiv:2104.13478), May 2021. doi: 10.48550/arXiv.2104.13478. URL http://arxiv.org/abs/2104.13478. arXiv:2104.13478 [cs, stat].

[6] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, October 2001. ISSN 0090-5364, 2168-8966. doi: 10.1214/aos/1013203451.

[7] Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. S2orc: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, page 4969–4983, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.447. URL https://www.aclweb.org/anthology/2020.acl-main.447.

[8] Jason Priem, Heather Piwowar, and Richard Orr. Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. 2022. doi: 10.48550/ARXIV.2205.01833. URL https://arxiv.org/abs/2205.01833.

[9] Stanisław Giziński, Paulina Kaczyńska, Hubert Ruczyński, Emilia Wiśnios, Bartosz Pieliński, Przemysław Biecek, and Julian Sienkiewicz. Big tech influence over ai research revisited: Memetic analysis of attribution of ideas to affiliation. *Journal of Informetrics*, 18(4):101572, November 2024. ISSN 17511577. doi: 10.1016/j.joi.2024.101572.

[10] Johannes C. Paetzold, Julian McGinnis, Suprosanna Shit, Ivan Ezhov, Paul Büschl, Chinmay Prabhakar, Mihail I. Todorov, Anjany Sekuboyina, Georgios Kaissis, Ali Ertürk, Stephan Günnemann, and Bjoern H. Menze. Whole brain vessel graphs: A dataset and benchmark for graph learning and neuroscience (vesselgraph), August 2021. URL http://arxiv.org/abs/2108.13233. arXiv:2108.13233 [cs, q-bio].

[11] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. 2016. doi: 10.48550/ARXIV.1609.02907. URL https://arxiv.org/abs/1609.02907.

[12] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. (arXiv:1710.10903), February 2018. doi: 10.48550/arXiv.1710.10903. URL http://arxiv.org/abs/1710.10903. arXiv:1710.10903 [cs, stat].

[13] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. (arXiv:1502.03167), March 2015. doi: 10.48550/arXiv.1502.03167. URL http://arxiv.org/abs/1502.03167. arXiv:1502.03167 [cs].

[14] Matthias Fey and Jan Eric Lenssen. Fast graph representation learning with pytorch geometric. 2019. doi: 10.48550/ARXIV.1903.02428. URL https://arxiv.org/abs/1903.02428.

[15] Kalina Hristova and William C. Wimley. Determining the statistical significance of the difference between arbitrary curves: A spreadsheet method. *PLOS ONE*, 18(10):e0289619, October 2023. ISSN 1932-6203. doi: 10.1371/journal.pone.0289619.

[16] Michael Färber and Lazaros Tampakis. Analyzing the impact of companies on ai research based on publications. *Scientometrics*, 129(1):31–63, January 2024. ISSN 0138-9130, 1588-2861. doi: 10.1007/s11192-023-04867-3.

A Algorithms

Algorithm 1: ALE Exact Version

Input: Model M, Dataset $\mathcal{D}$, Feature index f, Number of bins N
Output: Accumulated Local Effects (ALE) values

Initialize empty list ALE;
Divide feature values into N bins;
for *each bin* b_i **do**
  Get nodes in bin b_i;
  for *each node* n_j *in* b_i **do**
    `/* Additional loop in Exact version */`
    Set feature f of n_j to lower bin edge;
    Compute prediction P_{low};
    Set feature f of n_j to upper bin edge;
    Compute prediction P_{high};
    Compute difference $D = P_{high} - P_{low}$;
    Store D;
  Compute average difference for bin b_i and update ALE;
Return ALE;

Algorithm 2: ALE Approximate Version

Input: Model M, Dataset $\mathcal{D}$, Feature index f, Number of bins N
Output: Accumulated Local Effects (ALE) values

Initialize empty list ALE;
Divide feature values into N bins;
for *each bin* b_i **do**
  Get nodes in bin b_i;
  Set feature f of all nodes in b_i to lower bin edge;
  Compute prediction P_{low};
  Set feature f of all nodes in b_i to upper bin edge;
  Compute prediction P_{high};
  Compute average difference $D = P_{high} - P_{low}$;
  Update ALE with D;
Return ALE;

B Example of modified nodes' interaction

Figure 4 presents a hypothetical scenario in which modified nodes could affect each others' prediction produced by a two-layer GNN similar to the ones used in this article. Figure 5 is a modified version where the nodes would not affect each other embedding. All nodes inside of the dotted circle are connected with the node in the center of this circle by a path no longer than 2. Only information from the nodes inside the dotted circle can affect the embedding of the central node produced by a model with two layers of GCN or GAT. In the first layer, information from the node's neighbors is passed through the edges and aggregated. In the second layer, the same happens, but the neighbors embedding already contains information about neighbors' neighbors.

The more nodes are modified, the bigger the chance that some of them will be connected by a path short enough to influence each other. Hence, the disturbance coming from the interaction of modified nodes should rise with the number of modified nodes.

C Models

The models on the CD1-E_no2 dataset were trained on a GPU L40 with 24GB. Adam optimizer was used. The learning rate for GCN models was 10^{-6} and for GAT models was 10^{-5}.

D ALE profiles

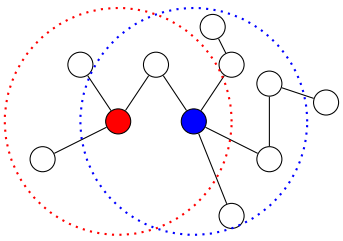

Figure 4: An exemplary graph where two modified features (blue and red) would affect each other during inference through the two-layer GNN.

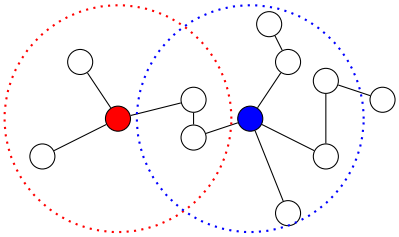

Figure 5: Modification of Fig. 4, where the second node was added on the path between the blue and red nodes. The blue and red nodes would no longer affect each other's embedding.

Model	Dataset	F1	AUC ROC
GAT	Citations	0.683	0.635
GCN	Citations	0.703	0.759
GAT	CD1-E_no2	0.741	–
GCN	CD1-E_no2	0.833	–

Table 1: Metrics for models trained on the Citation and CD1-E_no2 datasets. AUC ROC is only applicable to the Citation dataset.

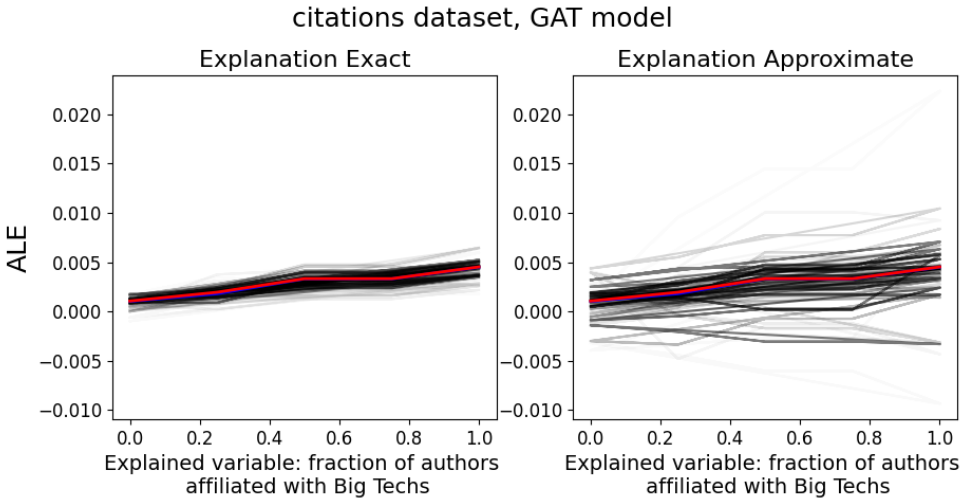

Figure 6: ALE curve calculated for the fraction of authors affiliated with Big Techs. The red line is the gold standard, and the blue line is the weighted average of approximate predictions.

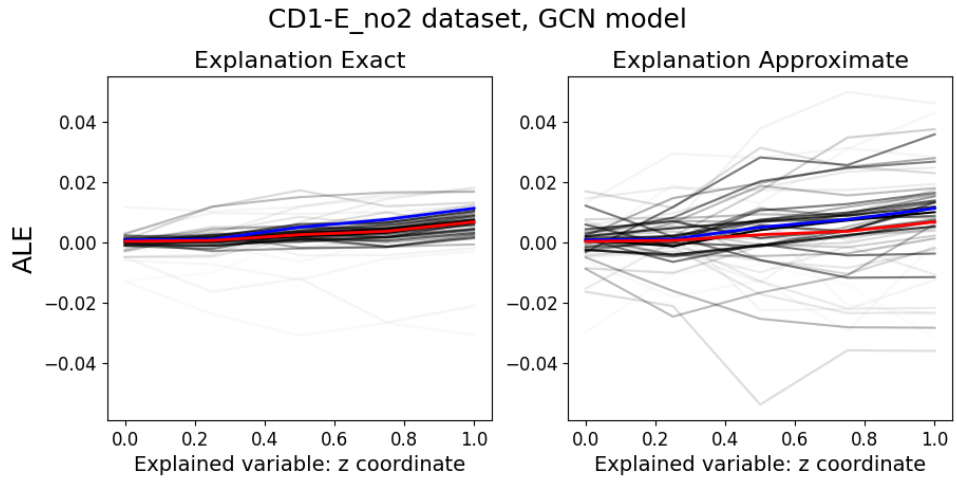

Figure 7: ALE curve calculated for the z coordinate in GCN model trained on CD1-E_no2 dataset. The red line is the gold standard, and the blue line is the weighted average of approximate predictions.

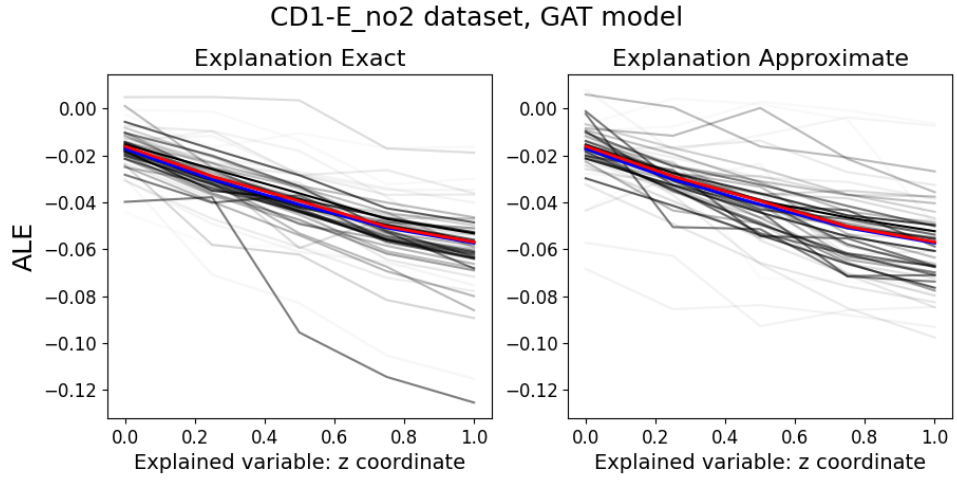

Figure 8: ALE curve calculated for the z coordinate in GAT model trained on CD1-E_no2 dataset. The red line is the gold standard and the blue line is the weighted average of approximate predictions.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We tried to accurately summarize in the abstract and introduction the key findings, objectives, and scope of the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper has Limitations section, which describes methodological challenges and underlines what was the scope of our work.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide details of the models' architecture, and the algorithm for approximate and exact ALE.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

The code is in the Github repository: `https://github.com/Kaczyniec/ALE-and-GNNs`.

Dataset CD1-E_no2 was made available on Github by its authors [10]. S2ORC database is in open access[7]. OpenAlex database is in open access as well [8].

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We tried to specify hyperparameters and other training details in the article and Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We performed statistical tests (χ^2 and permutation test) in order to test whether the two methods produce significantly different ALE profiles. We report results related to time and parameters without the error bars due to the choice of visualisation and exploratory nature of the considerations involving parameters.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: In case of CD1-E_no2 dataset we provided GPU details. Information about explanation time is available on the Fig. 2 and 3.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We read and followed the Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

Justification: We firmly believe that explainability tools like ALE can be important in diagnosing problems with ML models, which becomes increasingly critical as AI applications broaden in our daily lives. ALE differs from other methods used in GNN explainability and, in this way, complements them.

The choice of the Citations dataset was driven by our curiosity about Big Tech's impact on AI research, what has wide societal impacts.

We also include a subsection in the Discussion section highlighting how ALE can be interpreted. There, we emphasize that explainability tools describe models and do not necessarily represent real-life phenomena.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: We believe this study poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We strived to credit the owners of the assets and mention the authors of the assets were applicable.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: No assets (except code) were introduced. The code's README is in preparation. We believe that the experiments can be replicated with the current description of the code, but we are in the process of preparing more detailed instructions on how to run experiments.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: No crowdsourcing nor research with human subjects was performed.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: No research with human subjects was performed.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.