Vision Paper: Designing Graph Neural Networks in Compliance with the European Artificial Intelligence Act

Barbara Hoffmann
University of Bayreuth
Bayreuth, Germany
barbara.hoffmann@uni-bayreuth.de

Jana Vatter
Technical University of Munich
München, Germany
jana.vatter@tum.de

Ruben Mayer
University of Bayreuth
Bayreuth, Germany
ruben.mayer@uni-bayreuth.de

ABSTRACT

The European Union's Artificial Intelligence Act (AI Act) introduces comprehensive guidelines for the development and oversight of Artificial Intelligence (AI) and Machine Learning (ML) systems, with significant implications for Graph Neural Networks (GNNs). This paper addresses the unique challenges posed by the AI Act for GNNs, which operate on complex graph-structured data. The legislation's requirements for data management, data governance, robustness, human oversight, and privacy necessitate tailored strategies for GNNs. Our study explores the impact of these requirements on GNN training and proposes methods to ensure compliance. We provide an in-depth analysis of bias, robustness, explainability, and privacy in the context of GNNs, highlighting the need for fair sampling strategies and effective interpretability techniques. Our contributions fill the research gap by offering specific guidance for GNNs under the new legislative framework and identifying open questions and future research directions.

VLDB Workshop Reference Format:
Barbara Hoffmann, Jana Vatter, and Ruben Mayer. Vision Paper: Designing Graph Neural Networks in Compliance with the European Artificial Intelligence Act. VLDB 2024 Workshop: 3rd International Workshop on Large-Scale Graph Data Analytics (LSGDA 2024).

1 INTRODUCTION

The European Union has taken a significant step forward in the technological domain with the publication of the European Union Artificial Intelligence Act (AI Act) [41]. This legislation is notable for its provision of guidelines aimed at developing frameworks for Artificial Intelligence (AI) and overseeing Machine Learning (ML) practices. These frameworks possess considerable potential, as they could serve as a blueprint for future endeavors in the domain of AI and ML. The legislation introduces several requirements, for instance data management and data governance, robustness of training data and models and human oversight, which highly impacts Graph Neural Network (GNN) training. Therefore, it is important to know which requirements there are, understand their implications on GNN training and build the model accordingly.

GNNs have unique characteristics that warrant a closer examination. Unlike traditional ML models, GNNs operate on graph-structured data, which introduces complexities in data connectivity and relationships. These special characteristics justify the need for a detailed analysis of the AI Act's impact on GNNs. While there are initial studies examining the AI Act's effects on ML systems [24, 43, 45], no specific research has been conducted on GNNs. In this paper, we argue that GNNs pose specific challenges in terms of core requirements of the AI Act, such as data governance, robustness, explainability and privacy, that demand a closer investigation. This gap underscores the importance of our work in providing tailored guidance for GNNs under the new legislative framework and discussing open issues that demand further research.

In detail, our contributions are:

- Detailed examination of the requirements of the AI Act, tailored precisely to the use of GNNs to ensure compliance with the AI Act. This enhances the current discussion of the AI Act to the specifics of GNNs.
- Investigation of data and model bias in graphs and GNNs, particularly due to unfair sampling during GNN training. This opens a new perspective on data management techniques for GNNs beyond model accuracy and training runtime.
- Exploration of human oversight and explainability with concrete examples of GNN decision-making. We highlight the inherent trade-off between the comprehensibility and accuracy of an explanation, and articulate the demand for further (user) studies to better understand the effect of explanations on AI system stakeholders.
- Analysis of privacy-preserving techniques designed for GNN training. In particular, we stress that while adding noise to features and labels via Differential Privacy techniques is well-explored, GNNs additionally expose connections between entities—which can themselves contain privacy-sensitive information.
- Asking important questions arising from our investigations and identifying open research areas in this field. This promises to spawn further research in this area.

In Section 2, we provide an introduction to GNNs and explain the basics of the AI Act, detailing its implications for GNNs. In Section 3, we describe the experiments we conducted and present the results. The open questions that arise from these results are discussed in Section 4. Finally, we review related work in Section 5 and draw conclusions in Section 6.

2 BACKGROUND

This section gives a short introduction to GNNs as well as an overview of the four risk categories delineated by the AI Act, along with their implications for the training of GNNs. Additionally, an

Proceedings of the VLDB Endowment. ISSN 2150-8097.

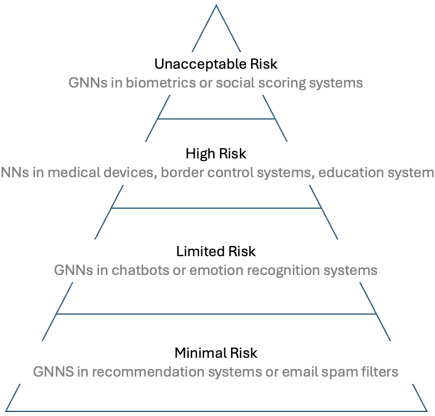

Unacceptable Risk
GNNs in biometrics or social scoring systems
High Risk
GNNs in medical devices, border control systems, education systems
Limited Risk
GNNs in chatbots or emotion recognition systems
Minimal Risk
GNNS in recommendation systems or email spam filters

Figure 1: Risk Pyramid as described in the AI Act

overview of the prerequisites outlined in the AI Act for high-risk systems is provided.

2.1 Graph Neural Networks

GNNs are specialized neural networks devised for analyzing data represented in graph forms [17, 36], such as networks found in social or citation systems. GNNs process data by transforming the initial node features into embeddings via an iterative mechanism known as message passing. During this process, each node computes new embeddings by aggregating and synthesizing the embeddings from its adjacent nodes, an operation underpinned by neural networks that are trainable at each layer of the GNN. The input features at the initial layer are raw node features, which are systematically enhanced through successive layers to refine the embeddings. This refinement process is driven by the objective of minimizing a predefined loss function, typically optimized through algorithms like stochastic gradient descent. The resulting embeddings, which encapsulate the essential structural and feature-based information of the nodes or the entire graph, are subsequently utilized in downstream applications such as node classification, link prediction, and graph classification. This makes GNNs particularly effective for tasks where data is inherently structured as graphs, such as social networks or citation networks. For a deeper technical explanation of GNNs, Vatter et al. [42] can be consulted.

2.2 AI Act Basics

The European Artificial Intelligence Act contains a classification schema for Artificial Intelligence systems. This schema is predicated on a risk-based approach and segregates AI systems into four distinct categories.

1. Minimal Risk: AI systems falling under this category are characterized by their low potential to inflict harm upon the user. Examples of such systems include GNNs utilized in decision support systems, such as recommendation systems [15, 46], or GNNs employed in spam filters for emails [11]. It is noteworthy that the regulations stipulated by the EU AI Act do not extend to this category of AI systems, thereby obviating the need for any regulatory action.
2. Limited risk: This category refers to risks associated with a lack of transparency in AI usage. AI systems that fall into this category are subject to the requirement of extended transparency. The category comprises for example AI systems that use GNNs in chatbots [29] or emotion recognition systems [26]. The user interacting with these kinds of AI systems needs to be made aware of interacting with a machine so that they can take an informed decision to continue or step back. The content generated by such systems must also be marked as AI generated.
3. High risk: Encompasses AI systems that process critical personal data or could put the life and health of citizens at risk. Examples of such systems include medical devices [16, 39], systems that control access to education [37], and border control systems. GNNs can be used in all the areas mentioned. For these systems, an enhanced level of quality management is mandated. The criteria for this include data governance, a possibility for human oversight and the robustness of the applications.
4. Unacceptable Risk: AI systems that fall into this category are deemed illegal. These are for example systems that manipulate the user or contribute to social injustice.

With Article 53, the AI Act also includes a category for General Purpose AI (GPAI). An AI system is classified as GPAI if it is designed to perform a wide variety of tasks across different contexts and applications. This encompasses capabilities like image and speech recognition, audio and video generation, pattern detection, question answering, and translation. GPAI systems are subject to additional documentation and risk management requirements, beyond the standard risk-based requirements. Furthermore, outlined in Article 51, there is a subcategory for Systemic Risk, which is determined by the computing power used during training.

2.3 AI Act Requirements for GNNs

The risk category denoted as "High Risk" warrants special attention in this context. The subsequent section provides an overview of the pivotal requirements delineated within the AI Act concerning the training and operation of high-risk GNNs. All of these criteria are outlined in the legal document [41], with a summary provided in Table 1 for quick reference.

Risk category	Article	Section
Data Governance	10	2.3.1
Robustness	15	2.3.2
Human Oversight	14	2.3.3
Privacy	(69), (27), (57), 5	2.3.4

Table 1: Overview: Requirements in the AI Act

2.3.1 Data Management and Data Governance.

The AI Act mandates that the data used in AI systems must comply with data governance requirements. This encompasses data sets utilized for training, validation, and testing purposes.

There is no universally applicable definition of the term *data governance*, neither in the scientific community nor among practitioners in the field of information systems [2, 32]. The definition provided by the AI Act involves ensuring that the data used for training, validation, and testing is relevant, representative, accurate, and as error-free as possible. This includes practices such as proper data collection and preparation, detecting and mitigating bias, and ensuring the data is suitable for the AI system's intended purpose.

In the context of the AI Act, Article 10 delineates the protocols for ensuring effective data governance. It necessitates a comprehensive scrutiny of AI systems, particularly concerning the potential for bias. The manifestation of bias, whether during the training phase or the deployment of AI systems, can precipitate detrimental effects. The primary objective is to forestall any bias that could compromise individual safety and health, infringe upon fundamental rights, or engender discrimination. As per the stipulations of the EU AI Act, such bias must be detected, prevented and mitigated. This is an essential prerequisite for maintaining the integrity and fairness of AI systems.

2.3.2 Robustness. According to the AI Act, Article 15, AI systems designed to continue learning post-deployment must be designed to reduce or eliminate the risk of biased outcomes. This requirement also includes ensuring that any potential biases are effectively addressed through suitable risk mitigation strategies. This process of ongoing learning and potential bias reinforcement is referred to as a *feedback loop*.

The robustness of a model refers to its ability to provide consistent and reliable predictions across different data sets and under different conditions. A robust model should also be able to respond well to new, unknown data or to data with minor disturbances. If a model is biased, it is less robust. Bias occurs when a model systematically prefers or excludes specific elements of the data, which may arise from imbalances in training data, errors in model design, or suboptimal sampling techniques.

2.3.3 Explainability for Humans. Article 14 of the AI Act stipulates that AI systems must be engineered to enable effective human oversight, ensuring the minimization of risks to health, safety, and fundamental rights. The legislation further mandates that the outputs of these AI systems should be interpretable and the derivation of the results should be understandable. Many AI models, in particular deep neural networks, are referred to as a black box. To make decisions comprehensible, existing interpretability techniques and methodologies from the domain of explainable artificial intelligence (xAI) can be utilized.

Common xAI methods include model-agnostic techniques like Local Interpretable Model Agnostic Explanations (LIME) [35], which approximates black-box models locally with interpretable ones, and SHapley Additive exPlanations (SHAP) [28], which assigns importance values to features based on cooperative game theory. Model-specific methods include feature importance for tree-based models [6] and visualization techniques like saliency maps for neural networks [38], all aimed at increasing the interpretability of AI models.

To enhance the interpretability of GNNs, methods like GNNExplainer [47] have been developed. GNNExplainer has been implemented in PyTorch Geometric [33] as well as in the Deep Graph Library (DGL) [8]. This method operates post-hoc, meaning it explains GNN decisions after they are made. It accomplishes this by analyzing subgraphs within the larger input graph. The output consists of a concise subgraph from the original input, along with a selection of node features deemed most influential in driving the model's predictions. These explanations are localized to individual instances, necessitating retraining for each new instance [47].

2.3.4 Privacy. The EU AI Act contains various passages for the secure handling of personal data, emphasizing its protection and confidentiality. For instance, legal regulations mandate enhanced safeguarding of data utilized in the creation of AI systems or in the mitigation of bias within these systems. In the context of AI systems, adherence to all relevant data protection laws, such as the General Data Protection Regulation (GDPR), is obligatory. This safeguarding can be accomplished through the implementation of anonymization and encryption techniques, which in GNNs can be done by the anonymization of the features or by adding noise to the graph, for example by adding or removing edges or nodes [25, 36].

3 ANALYSIS

In this section, we examine the areas of influence — Data Governance, Robustness, Explainability, and Privacy — identified in Section 2.3 with regard to their precise impact on GNNs. The focus hereby lies on bias in the data and the model, human oversight and privacy.

3.1 Data Governance

High-risk AI systems must be trained with data that meets certain standards and requirements. An important point here is an investigation into possible biases that could affect the health and safety of individuals, have a negative impact on fundamental rights or lead to discrimination prohibited by the AI Act, especially if the data outputs influence the inputs for future operations [41].

If training data exhibits an unbalanced feature or label distribution, this can lead to the aforementioned bias. Technically speaking, any dataset that shows an unequal distribution among its classes can be regarded as imbalanced. Yet, the prevalent view within the community is that imbalanced data specifically refers to datasets with substantial disparities between classes. This specific type of inequality is known as between-class imbalance. Another form of imbalance is the within-class imbalance, which focuses on the distribution of representative data for various subconcepts within a single class [21]. In addition to the class imbalance there is also a label imbalance. The Difference in Proportions of Labels (DPL) metric compares the proportion of observed outcomes with positive labels in one subgroup to the proportion in another subgroup within a training dataset [4].

In this paper we focus on between-class imbalance as well as label imbalance and measure these values according to existing implementations [3, 4]. The results are shown in Table 2. The following applies to both imbalance metrics: the closer to zero, the better the distribution of the data set; the further away from zero, the greater the imbalance. As can be seen in Table 2, some datasets yield low class and label imbalance, while others show high imbalance in classes, labels, or both.

Key Takeaway: Class and label imbalance are common issues across various graph datasets. It is important to monitor such imbalance, and, if appropriate, take corrective action.

Table 2: Overview of the datasets

Name	#Nodes	#Edges	Sensitive Attribute	Feature Size	Class Imbal.	Label Imbal.	Label
german	1,000	44,484	gender	27	-0.380	-0.298	high/low credit risk
recidivism	18,876	642,616	ethnos	18	0.013	-0.033	bail/no bail
credit	30,000	304,754	age	13	-0.821	-0.648	payment default/no default
pokec-n	66,569	1,100,663	region	266	-0.422	-0.021	working field
pokec-z	67,796	1,303,712	region	277	-0.297	-0.021	working field

3.2 Robustness

Sampling is a method to efficiently train GNNs on large-scale graphs. When performing sampling during GNN training, a subset of data points - such as nodes, edges, or subgraphs - is selected from the full graph. Before each training epoch, new samples are constructed. As numerous sampling strategies with different objectives exist, the choice of sampling method can significantly influence both the performance and the bias of the model. Inadequate or disproportionate sampling may result in the neglect of crucial segments of the graph, thereby distorting the overall representation of the data. This, in turn, affects the model's ability to generalize and perform accurately, underscoring the intricate link between sampling strategies and the robustness of machine learning models [27]. In the following, we explore whether sampling can influence the bias during training as only a selected subset of nodes and edges is used for training. This could lead to an under-representation of certain classes or labels. For our experiments, we use the datasets `german`, `recidivism`, `credit`, `pokec-n` and `pokec-z` as well as the sampling strategies Neighbor [19], VR-GCN [7], LABOR [5], and ShaDow [48]. We additionally include *no sampling* as a baseline. The 2-layer GCN and the sampling strategies are implemented with DGL. Our evaluation is based on the Area Under the Curve (AUC), Statistical Parity [13] and Equality of Opportunity [20]. While Statistical Parity measures how independent the predictions of a model are to a sensitive attribute, Equality of Opportunity denotes to which extent the predictions are performed equally well across all attributes. For both fairness metrics, lower values indicate a fairer model, while for AUC, higher values are better.

In Figure 2, we show the experimental results. Across all datasets and metrics, the values of Neighbor sampling and LABOR usually are close to the baseline (*no sampling*). For parity and equality, they sometimes even lead to better results than the baseline, with the exception of LABOR leading to worse equality on `credit`. VR-GCN, on the other hand, has a higher AUC score than the baseline and other strategies, but can result in higher values of parity and equality, especially when using the `german` or `credit` credit graph. The fourth sampling strategy, namely ShaDow, proves less suitable. A larger bias is induced compared to the other methods, especially for the `german`, `credit`, and `pokec-z` dataset, while the AUC is lower than the baseline.

Neighbor sampling chooses the nodes and edges at random which is beneficial for the model bias since all groups and attributes are treated equally. VR-GCN also is a node-wise method, but uses importance scores to prioritize certain nodes. As VR-GCN is based on historical activations, valuable information is preserved during the sampling step, but bias can be reinforced by favoring selected nodes. In this way, certain groups or attributes might be over- or underrepresented in the samples. LABOR is a layer-wise strategy using a specialized optimization method based on vertex-variance and restricting the size of the neighborhood to a small number. Therefore, a higher AUC can be achieved, but the model might not be as fair compared to other methods due to the specialized selection of nodes and edges. The fourth method, namely ShaDow, works in a subgraph-based fashion. The strategy first aims to form shallow subgraphs with a depth typically around 2 or 3. After that, sampling takes place within the subgraphs. Possibly important connections between and within clusters might be cut, which could lead to a loss of information needed for training and higher parity and equality values.

Key Takeaway: Our experiments have shown that bias can be induced or reinforced when using sampling-based GNN training. Some strategies lead to higher performance values, but also to a more biased model. Robustness against model bias needs to be taken into account when designing GNN sampling methods. This aspect has often been neglected.

3.3 Explainability

In this section, we delve deeper into the possibilities for GNN explanation. Our exploration is based on the two implementations of GNNExplainer outlined in Section 2.3.3 which allow for the visualization of subgraphs and feature importance.

GNNExplainer is capable of being applied to various scenarios, including graph classification and link prediction. In this paper, we restrict the scope of our experiments to node classification. For the explanations provided, a basic Graph Convolutional Network (GCN) consisting of two convolutional layers was utilized. This network integrates linear transformations with Rectified Linear Unit (ReLU) activation functions and incorporates dropout to enhance generalization. We used the dataset `german` as the foundational data for these experiments. The dataset offers insights into whether a customer with specific features qualifies as a good customer with low credit risk. Similarly, the node classification addresses the same question: determining whether the selected node, representing a customer, is credit-worthy or not.

The experiments were conducted using both 2-hop and 1-hop neighborhoods in PyTorch Geometric[1]. Figure 3 illustrates the explanations generated from these experiments. It is evident that

[1] The visualizations generated by DGL are consistent in content, despite differences in their layout. Due to this uniformity in content, they are not depicted in the paper.

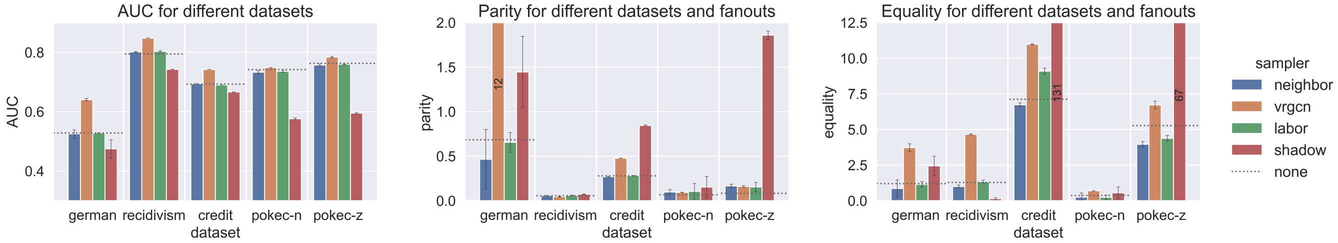


Figure 2: Results for the metrics AUC, parity, and equality for the different datasets and samplers. AUC: higher values are better (1.0 is best). Parity and equality: lower values are better (0.0 is best).

increasing the number of hops results in a more complex explanatory graph. Figure 3b shows the 2-hop result in a representation that is essentially imperceptible to the human eye, potentially making it challenging for many stakeholders of an AI system to comprehend. For the 1-hop neighborhood (Figure 3a), the graph remains simple and comprehensible for human interpretation; however, it raises the question of whether such simplicity adequately captures the complexity of node classification. In both scenarios, only the nodes that impacted the decision are depicted, and it remains uncertain whether this is sufficient for a meaningful explanation. For further analysis, nodes can be mapped to their corresponding features, of which an excerpt is shown in Table 4.

The features of the nodes are pivotal in the process of node classification. The visualization output of PyTorch Geometric is displayed in Figure 3c, which shows the most important features leading to a specific node's classification[2]. Additionally, the mapping of feature numbers to their corresponding meanings is detailed in Table 3.

Feature Number	Meaning
5	Loan Duration
3	Single
6	Purpose of Loan
4	Age
9	Years at current home
7	Loan Amount
8	Loan Rate as percent of income
21	Other Loans at store
2	Foreign worker
26	Unemployed

Table 3: Mapping of important features

Key Takeaway: Graphs are inherently complex structures. Consequently, methods that elucidate the behavior of a GNN through a graph also tend to be intricate. Simplified explanations, such as using only a 1-hop neighborhood or focusing solely on feature importance without including graph information, are more straightforward to comprehend but provide less detailed information. So there is an inherent trade-off between simplicity and the depth of information in GNN explanations.

3.4 Privacy

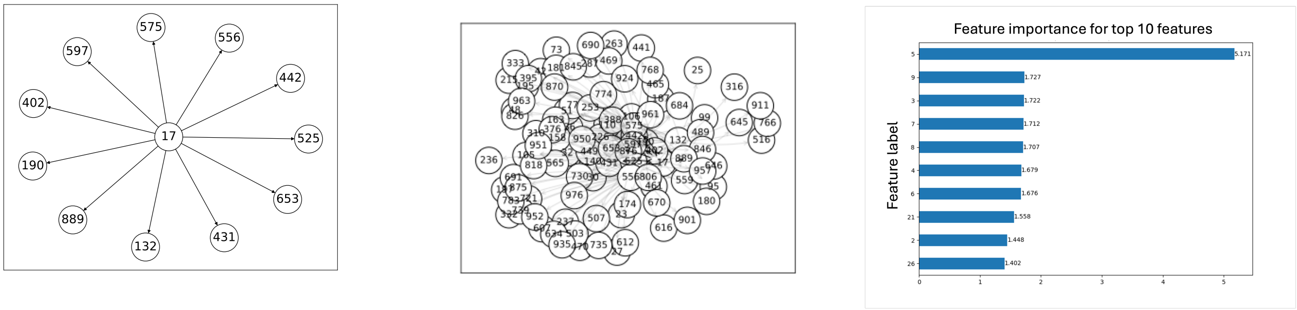
When it comes to maintaining privacy within graph-based data, one of the main concerns are the node features. Features must be kept confidential and anonymized using appropriate methods as needed. Before training a GNN, data anonymization techniques can be employed to inhibit any potential identification of individual users, thus protecting user privacy throughout the model training process. When GNNs are employed, they process edges using machine learning. This raises privacy concerns regarding the edges: Should they be considered private data associated with a specific node, or are they exempt from privacy considerations?

Assuming the data is confidential, Differential Privacy (DP) [12] could be employed as a strategy to protect privacy. The fundamental principle of DP is that when querying a dataset consisting of N individuals, the outcome should be, in probabilistic terms, virtually the same as if the query were run on a similar dataset that has either one fewer or one additional individual. This approach ensures the privacy of each individual with a certain probability. To achieve this level of probabilistic indistinguishability, adequate noise is added to the results of the query, masking individual data points while still providing useful aggregate information [31].

When applying DP to GNNs, several challenges arise. As outlined before, DP involves adding noise to the data, which, in the context of graphs, could mean adding or removing edges. Such modifications can significantly alter the dataset. If DP is implemented in a GNN, it is essential to evaluate whether the anonymity provided compromises the utility of the model. In cases of uncertainty, the importance of maintaining privacy versus the utility of the data must be carefully weighed. Moreover, it is crucial to determine the maximum amount of noise that can be introduced before the data loses its meaning due to excessive alteration. This threshold may vary from one GNN to another, as different structures have different tolerances for noise. However, these questions are largely unexplored.

Key Takeaway: Privacy in GNNs concerns not only the node features, but also the structural information of the graph itself, i.e., the edges. More research is needed on privacy-preserving GNN training and inference.

[2] An alternative method to explain GNNs is through the GraphLIME framework [22]. GraphLIME utilizes an Hilbert-Schmidt Independence Criterion (HSIC) Lasso model to provide localized, nonlinear explanations for the predictions made by GNNs. These explanations are limited to the K most representative features as the explanation for the prediction of a particular node. The approach integrates both the local structure of the graph and nonlinear dependencies to enhance understanding. Despite employing a distinct approach, GraphLIME solely focuses on visualizing the feature importance and the output looks similar to Figure 3c.



(a) Subgraph 1-hop

(b) Subgraph 2-hop

(c) Important features

Figure 3: Overview of PyTorch Geometric generated grahps and important features

Table 4: Listing of important nodes with a selection of their features

	Credit Worthy	Gender	Foreign Worker	Single	Age	Loan Duration	PurposeOf Loan	Loan Amount	LoanRateAsPer-centOfIncome
597	-1	Male	0	1	36	24	Business	4241	1
190	-1	Male	0	1	54	24	Business	4591	2
556	-1	Female	0	0	28	18	NewCar	2278	3
439	-1	Female	0	0	26	12	Business	609	4
575	1	Female	0	0	24	15	Furniture	2788	2
132	1	Male	0	1	27	15	Furniture	2708	2
442	1	Male	0	1	29	20	Other	2629	2
889	1	Male	0	1	40	28	UsedCar	7824	3
525	1	Male	0	1	30	26	UsedCar	7966	2
402	-1	Male	0	1	27	24	Business	8648	2
653	-1	Male	0	1	42	36	NewCar	8086	2

4 OPEN QUESTIONS AND RESEARCH DIRECTIONS

In this section, we look at the questions that arise from our investigations and experiments and propose new research directions.

4.1 Bias and Robustness

For our experiments, we use standard sampling methods which do not specifically aim at reducing bias. Our results show that fairness can highly depend on the chosen sampling method. Consequently, the question arises how to better ensure fairness during sampling. Fair random walk strategies, such as those proposed by Rahman et al. [34] and Zhang et al. [49] could be considered in GNN sampling. Further, our experiments are evaluated with metrics commonly used in the field of machine learning. However, when using graphs and GNNs, other factors such as feature distribution and structure, particularly the nature of the connections, play a crucial role in regards to fairness. More research is needed in the directions of designing fairness metrics adapted to the specific characteristics of GNNs.

Furthermore, it is essential to consider whether GNNs are robust against data distribution shifts. A data distribution shift occurs when the data a model uses changes over time, leading to a decline in prediction accuracy. Ensuring robustness in GNNs means maintaining accurate classification performance with new and evolving data.

We summarize the open questions as follows: How can sampling strategies be specifically adapted and optimized for different types of graph data to ensure comprehensive fairness without compromising model performance? What additional or refined fairness metrics need to be developed? What specific adaptations are needed for GNNs to handle dynamic and evolving graph data, and how can these techniques be seamlessly integrated into GNN frameworks?

4.2 Explainability and Privacy

Several questions arise in the context of explainability. Firstly, for whom the output of the AI system needs to be explained. Various stakeholders could come into consideration: The end customer, who may be influenced by a system's decision, the employee of the company in which the AI system is used and the auditor of a supervisory authority, who examines compliance with the AI Act in companies, could all be equally interested.

Each of these parties has a unique perspective on the required explanation of an AI decision. For example, the developer of an AI system seeks highly precise explanations to understand its functionality and requires specific information for debugging. They typically prioritize less on protecting the privacy of individual users' information, viewing the data as abstract rather than personal. Conversely, for the user, protecting personal data is a top priority, and the accuracy of the explanation may be secondary to this concern. For auditors, the primary interest is ensuring that there is an explanation of the AI system's decisions. Likewise, each of these individuals has their own level of authorization for insights. This implies several privacy issues, as it must be clarified who has the right to view specific data. To illustrate this concept with a specific example: In an online social network, User A receives the explanation that Group G was suggested to him because contact B and his contact C - who is not directly connected to A - are also in similar groups. This explanation allows User A to gain insight into contact C's affiliations, even though C has not directly shared this information with A.

It is also important to evaluate whether the selected explanatory approaches are appropriate. While explanations using subgraphs and key features are commonly employed, alternative forms such as text or images might be more comprehensible to some groups of users. Additionally, the methods we analyzed only highlight the important nodes and features. However, having an overview of the *unimportant* ones might also be beneficial, as it could provide insights for improving the training process of an AI model. To make an informed assessment of which method is preferred and which information in detail would be helpful for different user groups, conducting a user study would be essential [23]. Finally, assessing the effectiveness of explanability methods for human oversight is an interdisciplinary effort that requires further research [40], especially for GNNs which are potentially much more difficult to explain and understand due to their graph structure.

5 RELATED WORK

Various papers have already focused on presenting the content of the AI Act in an understandable way [14, 30] or explained the impact of the AI Act on ML systems [24, 43, 45]. The sub-topics we have identified as challenges for GNN training have also been highlighted in the literature. For example, [9] and [10] address bias in GNNs. As mentioned in Section 2.3.3, several methods exist for making GNNs explainable, including GNNExplainer [47] and GraphLiME [22], which we use in our work. Other approaches focus on more theoretical aspectes of explanations [1]. Studies have also tackled the security and privacy of GNNs, either by explaining security vulnerabilities and privacy-enhancing measures, as seen in survey articles [18, 50] or practical implementations such as SecGNN [44].

Despite these contributions, the intersection of the two topics - how the requirements set out specifically in the AI Act affect GNNs - has not yet been investigated. Our work uniquely fills this gap by offering a detailed examination of how GNNs can be designed to comply with the AI Act. By addressing these points, our work uniquely contributes to the field by bridging the gap between GNN technology and regulatory compliance, providing practical guidance and highlighting the need for further research on the intersection of these topics.

6 CONCLUSION

The AI Act establishes significant legal requirements for AI and ML, impacting crucial areas within these fields. In our paper, we demonstrated that the AI Act also presents critical challenges for GNNs. Our initial research on this topic has uncovered numerous additional open questions that need to be addressed.

To advance this field, we propose several concrete steps for future research. One key area is developing compliance rules specifically tailored to GNNs, including initial hypotheses on integrating AI Act aspects into GNN design and training. Additionally, more detailed studies on bias mitigation techniques within GNNs are necessary, focusing on models that reduce bias while maintaining performance. Enhancing explainability methods, such as refining GNNExplainer and GraphLiME, is also crucial for balancing transparency and accuracy. Finally, privacy-preserving techniques for GNNs warrant further exploration. Future research should focus on implementing methods like Differential Privacy and investigating their impact on privacy and model efficacy. By addressing these areas, we aim to bridge the gap between regulatory compliance and technological advancement in GNNs.

REFERENCES

[1] Chirag Agarwal, Marinka Zitnik, and Himabindu Lakkaraju. 2022. Probing gnn explainers: A rigorous theoretical and empirical analysis of gnn explanation methods. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 8969–8996.

[2] Majid Al-Ruithe, Elhadj Benkhelifa, and Khawar Hameed. 2019. A systematic literature review of data governance and cloud data governance. *Personal and Ubiquitous Computing* 23 (2019), 839–859.

[3] AWS. 2024. Class Imbalance. Retrieved June 18, 2024 from https://docs.aws.amazon.com/sagemaker/latest/dg/clarify-bias-metric-class-imbalance.html

[4] AWS. 2024. Difference in Proportions of Labels. Retrieved June 18, 2024 from https://docs.aws.amazon.com/sagemaker/latest/dg/clarify-data-bias-metric-true-label-imbalance.html

[5] Muhammed Fatih Balin and Ümit Çatalyürek. 2024. Layer-Neighbor Sampling—Defusing Neighborhood Explosion in GNNs. *Advances in Neural Information Processing Systems* 36 (2024).

[6] Leo Breiman. 2001. Random forests. *Machine learning* 45 (2001), 5–32.

[7] Jianfei Chen, Jun Zhu, and Le Song. 2018. Stochastic Training of Graph Convolutional Networks with Variance Reduction. In *International Conference on Machine Learning*. PMLR, 942–950.

[8] DGL Team. 2024. GNNExplainer. Retrieved June 8, 2024 from https://docs.dgl.ai/en/0.8.x/generated/dgl.nn.pytorch.explain.GNNExplainer.html

[9] Yushun Dong, Ninghao Liu, Brian Jalaian, and Jundong Li. 2022. Edits: Modeling and mitigating data bias for graph neural networks. In *Proceedings of the ACM web conference 2022*. 1259–1269.

[10] Yushun Dong, Song Wang, Yu Wang, Tyler Derr, and Jundong Li. 2022. On structural explanation of bias in graph neural networks. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 316–326.

[11] Tiny du Toit and Hennie Kruger. 2012. Filtering spam e-mail with generalized additive neural networks. In *2012 Information Security for South Africa*. IEEE, 1–8.

[12] Cynthia Dwork. 2006. Differential privacy. In *International colloquium on automata, languages, and programming*. Springer, 1–12.

[13] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. 214–226.

[14] Lilian Edwards. 2021. The EU AI Act: a summary of its significance and scope. *Artificial Intelligence (the EU AI Act)* 1 (2021).

[15] Chen Gao, Yu Zheng, Nian Li, Yinfeng Li, Yingrong Qin, Jinghua Piao, Yuhan Quan, Jianxin Chang, Depeng Jin, Xiangnan He, et al. 2023. A survey of graph neural networks for recommender systems: Challenges, methods, and directions. *ACM Transactions on Recommender Systems* 1, 1 (2023), 1–51.

[16] Grigoriy Gogoshin and Andrei S Rodin. 2023. Graph neural networks in cancer and oncology research: Emerging and future trends. *Cancers* 15, 24 (2023), 5858.

[17] Marco Gori, Gabriele Monfardini, and Franco Scarselli. 2005. A new model for learning in graph domains. In *Proceedings. 2005 IEEE Int. Joint Conf. on Neural Networks, 2005.*, Vol. 2. IEEE, 729–734.

[18] Faqian Guan, Tianqing Zhu, Wanlei Zhou, and Kim-Kwang Raymond Choo. 2024. Graph neural networks: a survey on the links between privacy and security. *Artificial Intelligence Review* 57, 2 (2024), 40.

[19] Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. *Advances in neural information processing systems* 30 (2017).

[20] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems* 29 (2016).

[21] Haibo He and Edwardo A Garcia. 2009. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering* 21, 9 (2009), 1263–1284.

[22] Qiang Huang, Makoto Yamada, Yuan Tian, Dinesh Singh, and Yi Chang. 2022. Graphlime: Local interpretable model explanations for graph neural networks. *IEEE Transactions on Knowledge and Data Engineering* (2022).

[23] Markus Langer, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesing, and Kevin Baum. 2021. What do we want from Explainable Artificial Intelligence (XAI)? – A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence* 296 (2021), 103473. https://doi.org/10.1016/j.artint.2021.103473

[24] Johann Laux, Sandra Wachter, and Brent Mittelstadt. 2024. Trustworthy artificial intelligence and the European Union AI act: On the conflation of trustworthiness and acceptability of risk. *Regulation & Governance* 18, 1 (2024), 3–32.

[25] Wen-Yu Lee. 2022. Graph Convolutional Networks Using Node Addition and Edge Reweighting. In *International Symposium on Methodologies for Intelligent Systems*. Springer, 368–377.

[26] Chenyu Liu, Xinliang Zhou, Yihao Wu, Ruizhi Yang, Liming Zhai, Ziyu Jia, and Yang Liu. 2024. Graph Neural Networks in EEG-based Emotion Recognition: A Survey. *arXiv preprint arXiv:2402.01138* (2024).

[27] Xin Liu, Mingyu Yan, Lei Deng, Guoqi Li, Xiaochun Ye, and Dongrui Fan. 2021. Sampling methods for efficient training of graph convolutional networks: A survey. *IEEE/CAA Journal of Automatica Sinica* 9, 2 (2021), 205–234.

[28] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30 (2017).

[29] Vitaly M Makarkin, Olesya V Matukhina, and Robert G Mukharlyamov. 2019. Development of a Neural Network Chatbot for User Support. In *The International Scientific and Practical Forum "Industry. Science. Competence. Integration"*. Springer, 624–630.

[30] Sean Musch, Michael Borrelli, and Charles Kerrigan. 2023. The EU AI Act As Global Artificial Intelligence Regulation. *Available at SSRN 4549261* (2023).

[31] Iyiola E Olatunji, Thorben Funke, and Megha Khosla. 2021. Releasing graph neural networks with differential privacy guarantees. *arXiv preprint arXiv:2109.08907* (2021).

[32] Boris Otto and Kristin Weber. 2011. Data governance. *Daten-und Informationsqualität: Auf dem weg zur information excellence* (2011), 277–295.

[33] PyG Team. 2024. torch_geometric.explain. Retrieved June 5, 2024 from https://pytorch-geometric.readthedocs.io/en/2.5.3/modules/explain.html#explainer

[34] Tahleen Rahman, Bartlomiej Surma, Michael Backes, and Yang Zhang. 2019. Fairwalk: Towards Fair Graph Embedding. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*. 3289–3295. https://doi.org/10.24963/ijcai.2019/456

[35] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.

[36] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. 2008. The graph neural network model. *IEEE transactions on neural networks* 20, 1 (2008), 61–80.

[37] Pengyang Shao, Chen Gao, Lei Chen, Yonghui Yang, Kun Zhang, and Meng Wang. 2024. Improving Cognitive Diagnosis Models with Adaptive Relational Graph Neural Networks. *arXiv preprint arXiv:2403.05559* (2024).

[38] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034* (2013).

[39] Aryan Singh, Pepijn Van de Ven, Ciarán Eising, and Patrick Denny. 2023. Compact & Capable: Harnessing Graph Neural Networks and Edge Convolution for Medical Image Classification. *arXiv preprint arXiv:2307.12790* (2023).

[40] Sarah Sterz, Kevin Baum, Sebastian Biewer, Holger Hermanns, Anne Lauber-Rönsberg, Philip Meinel, and Markus Langer. 2024. On the Quest for Effectiveness in Human Oversight: Interdisciplinary Perspectives. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) *(FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 2495–2507. https://doi.org/10.1145/3630106.3659051

[41] THE EUROPEAN PARLIAMENT AND THE COUNCIL OF THE EUROPEAN UNION. 2024. Artificial Intelligence Act. *European Parliament* (2024). https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.pdf

[42] Jana Vatter, Ruben Mayer, and Hans-Arno Jacobsen. 2023. The evolution of distributed systems for graph neural networks and their origin in graph processing and deep learning: A survey. *Comput. Surveys* 56, 1 (2023), 1–37.

[43] Michael Veale and Frederik Zuiderveen Borgesius. 2021. Demystifying the Draft EU Artificial Intelligence Act—Analysing the good, the bad, and the unclear elements of the proposed approach. *Computer Law Review International* 22, 4 (2021), 97–112.

[44] Songlei Wang, Yifeng Zheng, and Xiaohua Jia. 2023. Secgnn: Privacy-preserving graph neural network training and inference as a cloud service. *IEEE Transactions on Services Computing* (2023).

[45] Herbert Woisetschläger, Alexander Erben, Bill Marino, Shiqiang Wang, Nicholas D Lane, Ruben Mayer, and Hans-Arno Jacobsen. 2024. Federated Learning Priorities Under the European Union Artificial Intelligence Act. *arXiv preprint arXiv:2402.05968* (2024).

[46] Shiwen Wu, Fei Sun, Wentao Zhang, Xu Xie, and Bin Cui. 2022. Graph neural networks in recommender systems: a survey. *Comput. Surveys* 55, 5 (2022), 1–37.

[47] Rex Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. 2019. Gnn explainer: A tool for post-hoc explanation of graph neural networks. *arXiv preprint arXiv:1903.03894* 8 (2019).

[48] Hanqing Zeng, Muhan Zhang, Yinglong Xia, Ajitesh Srivastava, Andrey Malevich, Rajgopal Kannan, Viktor Prasanna, Long Jin, and Ren Chen. 2021. Decoupling the depth and scope of graph neural networks. *Advances in Neural Information Processing Systems* 34 (2021), 19665–19679.

[49] Guixian Zhang, Debo Cheng, and Shichao Zhang. 2023. Fpgnn: Fair path graph neural network for mitigating discrimination. *World Wide Web* 26, 5 (2023), 3119–3136.

[50] Yi Zhang, Yuying Zhao, Zhaoqing Li, Xueqi Cheng, Yu Wang, Olivera Kotevska, Philip S Yu, and Tyler Derr. 2023. A Survey on Privacy in Graph Neural Networks: Attacks, Preservation, and Applications. *arXiv preprint arXiv:2308.16375* (2023).