

A Review of Methods for Explainable and Interpretable Graph Neural Networks

Art Tay

1 Introduction

Graphs provide an incredibly flexible structure for modeling complex data. Data can naturally appear as graphs, like molecules. We can reduce data to a graph, such as the key points of a image. We can even use graphs to add structure, such as grammatical relationships. Graph Neural Networks (GNNs) have become a popular choice for prediction and inference on graph data. At their core, GNNs work by iteratively updating node embeddings based on information from neighboring nodes. The idea is to use the graph’s structure to engineer better features. This message passing scheme allows GNNs to capture complex dependencies and patterns present within the graph structure. GNN architectures typically consist of multiple layers, each performing message passing and aggregation operations to refine the embeddings. These layers are often followed by pooling and dense prediction layers to produce the final output. Some important applications of graph classification include predicting chemical toxicity ([Bai et al. 2019](#)), classifying proteins ([Gallicchio and Micheli 2019](#)), and even detecting cancer from pathology slides ([Xiao et al. 2023](#)). While GNNs achieve remarkable predictive power, their complexity prevents the exaction of the scientific rationale. Like many deep learning models, the black box nature of GNNs prevents wide adoption. Without strong methods for understanding their predictions, GNNs are more susceptible to adversarial attacks and undetected discrimination. Inferential methods also serve to direct modeling efforts by highlighting common structures or features that may be predictive in other applications. These reasons underscore the critical importance of methods for explaining and interpreting GNN models.

Although explainability and interpretability are sometimes used interchangeably in the literature, we will adopt the distinction expressed in Yuan et al. (2022). In this article the authors say that a model is explainable if the models predictions can be reasoned post-hoc (Yuan et al. 2022). For example, the effect of a specific input variable can be estimated by either dropping or randomly permuting the variable and accessing the effect on the output (Breiman 2001). This paradigm could be directly applied to GNNs; however, researchers are often not interested in the statistical significance of tabular node and edge-level features. Most scientific questions focus on the importance of specific graph-level substructures. The challenge with GNN explanations is that, in addition to computing the importance of features, the high-level features themselves must first be identified. On the other hand, a model is interpretable if the model’s decision process can be readily understood by humans (Yuan et al. 2022). For example, a linear regression model is interpretable because each coefficient clearly defines the relationship between any input and output. An interpretable model is typically superior to an explainable one because similar to a partial F-test, explanation methods might identify that certain variables enhance the model, but they cannot explain how these variables contribute. On the flip side, interpretable models may not reach the performance levels of black-box models. GNN models are generally not interpretable because the relation expressed by the parameters tends to exceed human understanding. A direct translation from traditional statistics would be a circuit type analysis, which has work for image processing convolution neural networks (Olah et al. 2020). For graphs, this would involve using coefficients on subgraphs to produce predictions. Similar to the difficulties of explanation methods, these subgraphs would have to be identified, which can be computational complex and expensive given the combinatorial nature of graphs. Overcoming these challenges has substantial scientific impacts. In applications where GNNs demonstrate strong predictive power, it enables the formulation of testable scientific hypotheses about the nature of the classification. Conversely, in cases where GNNs exhibit weak predictive power, it helps identify and understand potential misunderstandings within the model.

1.1 General Notation

- Any graph G can be describe by X, A, E . The node feature, matrix, edge feature matrix, and adjacency matrix respectively.

- Let $X = [X_c, X_d]$, where X_c is the subset of continuous node features and X_d is the subset of one-hot discrete node features.
- Let $E = [E_c, E_d]$, denoted in the same manner.
- Let n represent the number of nodes in the graph and v represent the number of edges.
- For any graph, let ν denote the set of node and \mathcal{E} denote the set of edges.
- Let $\text{feat}_{(\cdot)}$ denote the number of features or columns in the the corresponding feature matrix.
- A is a binary $n \times n$ matrix where $A[i, j] = 1$ indicates that an edge exists between nodes labeled i and j .
- Let $\text{explaineer}(G; \Omega) = h_G^{(1)}, \dots, h_G^{(L)}, \rho_G$ be an L layer GNN model with parameters Ω that we would like to explain.
- Let \hat{Y}_G be the predicted class label for graph G predicted from $\text{explaineer}()$.
- Unless otherwise noted, assume any GNN model discussed is a graph-level classification model.

2 Analysis of Core Papers

2.1 GNNInterpreter (Wang and Shen 2024)

GNNInterpreter (Wang and Shen 2024) is a method for generating model-level explanations of GNN graph classification models. In general, explanation methods serve to elucidate which features within the data influence disparate predictions. These methods typically fall into two categories: instance-level and model-level. Instance-level explanations aim to unveil the model’s rationale behind a particular prediction. In domains such as image and text analysis, a prevalent approach involves masking or perturbing the instance and assessing the impact on the model’s prediction. On the other hand, model-level explanations seek to understand how a model generally distinguishes between classes. In image and text analysis, for instance, one common technique involves treating the input as a trainable parameter and optimizing the model’s prediction towards a specific class. Consequently, the resulting optimized input comprises a set of features strongly associated with the targeted class. GNNInterpreter provides model level explanations for GNN

in this manner. Formally, GNNInterpreter tries to learn the graph generating distribution for each class. GNNInterpreter works by optimizing the parameters of a generic graph generating distribution to produce samples that closely match the explaine’s understanding of the targeted class.

Graph generating distributions are hard to specify because there can be discrete and continuous elements of X , E and A . Furthermore, the interactions between these matrices can be complex. The authors tackle these issues by making two simplify assumptions. First, they assume that every graph is a *Gilbert random graph* (Gilbert 1959), where every possible edge as an independent fixed probability of occurring. Second, the author assume that the features of every node and edge are independently distributed. The justification of the first assumption is that the other common types of random graphs are not suitable for this application. Erdo-Renyi random graphs (ERDdS and R&wi 1959) have a fixed number of edges, which limits the diversity of explanations, Rado random graphs (Rado 1964) are infinite in size, and the random dot-product graph model is just a generalization of Gilbert’s model. The second assumption is justified by the fact that the parameters of the independent distributions will be updated jointly using the *explaine* model. Therefore, the *explaine*’s understanding of the latent correlation structure should be contained in the final estimates.

Although the graphs discuss in this paper only contain discrete features, X_c and E_c can be sampled from any continuous distribution that can be expressed as a location-scale family. Separating the stochastic and systematic components is necessary for gradient based optimization. It is commonly known as the *reparametrization trick*. The discrete feature matrices, (X_d, E_d, A) , need to be sampled from a continuous distribution for gradient based optimization, but the distribution has to have sampling properties close to a discrete distribution. The author assume that the true underlying distribution for every discrete node and edge feature is *categorical*. The categorical distribution is also know as the multi-bernoulli, where every sample has a fixed probability of being in one of the discrete categories. Let θ_{cat} represent the associated vector of un-normalized or relative probabilities, where each entries is > 0 . Then,

$$\operatorname{argmax} \log \theta_{\text{cat}} + \text{Gumbel}(0, 1) \sim \text{Cat}(\theta_{\text{cat}}) \quad (1)$$

The intuition is that the Gumbel distribution, which is the density of the maximum order statistic of i.i.d. standard normals, makes it a good candidate for modeling the winning or maximum probability category. Adding Gumbel noise to the logits should maintain the true relative proportions, but enough skewness such that every category has some probability of having the maximum noised logit. A proof of equation 1 is provide in section Section 4.3. Approximating the argmax function with the Softmax function allows for approximate categorical sampling that is differentiable w.r.t. θ_{cat} . The associate inverse uniform CDF sampling formula below is referred to as the *Gumbel-Softmax trick* or the *concrete distribution* (Maddison, Mnih, and Teh 2017).

$$\text{Softmax} \left(\frac{\log \theta_{\text{Cat}} - \log(-\log \epsilon)}{\tau} \right), \quad \epsilon \sim U[0, 1]. \quad (2)$$

τ is a hyperparameter that controls the degree of relaxation. Smaller value of τ approximate the discrete sampling better, but can result in numerical issues. The adjacency matrix can be sampled in a similar manner since the Bernoulli is just a special case of the categorical.

$$\text{sigmoid} \left(\frac{\log(\theta_A/(1 - \theta_A)) + \log \epsilon - \log(1 - \log \epsilon)}{\tau} \right), \quad \epsilon \sim U[0, 1], \quad (3)$$

where θ_A is an $n \times n$ matrix with each $[i, j]$ entry representing the relative probability that the ij edge exists. Equation 3 samples from the binary concrete distribution (Maddison, Mnih, and Teh 2017). Taken together, let

$$G_{\text{gen}} \sim \text{gen}(\Theta)$$

notate the combined graph generating distribution, where Θ is the set of all parameters from the independently sampled distributions.

An obvious objective is to maximize the likelihood that the *explainee* model predicts a sampled graph to be a member of the target class. Let $\tilde{\rho}$ denote the desired predicted probability vector. Then the above objective can be expressed as:

$$\mathcal{L}_{\text{pred}}(\Theta \mid G_{\text{gen}}) = \mathbb{E}_{G_{\text{gen}}} \text{CrtEnt}(\text{explainee}(G_{\text{gen}}), \tilde{\rho}) \quad (4)$$

While the above objective enforces a desirable property, it fails to be restrictive enough to generate realistic graphs. This is because the final prediction, $\rho_{G_{\text{gen}}}$ is compute using only final embeddings, $h_{G_{\text{gen}}}^{(L)}$. Normally $h_{G_{\text{gen}}}^{(L)}$ contains all the needed information from the graphs structure; however, the generation scheme allows the feature distribution to be optimized directly. This means that an explanation can ignore the graph structure and optimize towards the desired final embeddings. It is dangerously common for nonsensical features and structures at the graph level to end up on the desired side of the decision boundary. Empirically, the authors found that even completely random graphs can produce confidence and consistent predictions.

Dataset [Method]	Predicted Class Probability by GNN		Training Time Per Class
Is_Acyclic [XGNN]	Cyclic 0.076 ± 0.000	Acyclic 0.927 ± 0.000	45 s
Is_Acyclic [Ours]	Cyclic 0.999 ± 0.001	Acyclic 1.000 ± 0.000	20 s
Is_Acyclic [Random]	Cyclic 0.143 ± 0.155	Acyclic 0.857 ± 0.155	-
MUTAG [XGNN]	Mutagen 0.986 ± 0.057	Nonmutagen 0.991 ± 0.083	128 s
MUTAG [Ours]	Mutagen 1.000 ± 0.000	Nonmutagen 1.000 ± 0.000	12 s
MUTAG [Random]	Mutagen 0.068 ± 0.251	Nonmutagen 0.932 ± 0.251	-
ColorConsistency [Ours]	Consistent 0.968 ± 0.110	Inconsistent 1.000 ± 0.000	58 s
ColorConsistency [Random]	Consistent 0.017 ± 0.118	Inconsistent 0.983 ± 0.118	-

Figure 1: An abridged copy of Table 5 from Wang and Shen (2024).

For example, Figure 1 shows that random graphs had an average predicted probability of being non-mutagenic in the MUTAG dataset (Debnath et al. 1991) if 93.2%. In order to mitigate this issue, the authors proposed additional minimizing the cosine distance between the average embedding of all the observed graph from the targeted class, $\bar{h}_{G_c}^{(L)}$, and the embedding of the generated explanation:

$$\mathcal{L}_{\text{embed}}(\Theta \mid G_{\text{gen}}) = \mathbb{E}_{G_{\text{gen}}} \text{CosDist} \left(\bar{h}_{G_c}^{(L)}, h_{G_{\text{gen}}}^{(L)} \right). \quad (5)$$

2.2 D4Explainer (Chen et al. 2023)

2.3 ProtGNN (Zhang et al. 2021)

3 Synthesis of Core Papers

4 Technical Details

4.1 Methodology

4.2 Results

4.3 Proofs

Proof of Equation 1: WLOG assume we want to sample for a categorical distribution $\text{cat}(\theta_{\text{cat}})$ with a set of W categories each with a probability

$$\pi_{\omega} = \frac{\theta_{\text{cat}}[\omega]}{\sum_{i \in W} \theta_{\text{cat}}[i]}.$$

Define random variable

$$D = \operatorname{argmax}_{i \in W} \log \theta_{\text{cat}} + \text{Gumbel}(0, 1) = \operatorname{argmax}_{i \in W} \log \theta_{\text{cat}}[i] + \text{Gumbel}(0, 1)[i],$$

here $\text{Gumbel}(0, 1)[i]$ denotes the i^{th} i.i.d. sample from the associated Gumbel. In order for D to be a true categorical random variable, $\Pr[D = \omega]$ need to be π_{ω} . $D = \omega$ if and only if $\log \theta_{\text{cat}}[\omega] + \text{Gumbel}(0, 1)[\omega] > \log \theta_{\text{cat}}[i] + \text{Gumbel}(0, 1)[i] \ \forall i \in W \setminus \omega$. Now, let M_i denote a random variable

that follows a $\text{Gumbel}(\log \theta_{\text{cat}}[i], 1)$ distribution. Then,

$$\begin{aligned}
Pr[D = \omega] &= \mathbb{E}_{M_\omega} \prod_{i \in W \setminus \omega} Pr(M_i < m_\omega) \text{ i.i.d location shifted Gumbel distributions.} \\
&= \mathbb{E}_{M_\omega} \prod_{i \in W \setminus \omega} \exp(-e^{\log \theta_{\text{cat}}[i] - m_\omega}) \text{ Gumbel CDF.} \\
&= \mathbb{E}_{M_\omega} \exp\left(-\sum_{i \in W \setminus \omega} e^{\log \theta_{\text{cat}}[i] - m_\omega}\right) \\
&= \int_{-\infty}^{\infty} \exp(\log \theta_{\text{cat}}[\omega] - m_\omega) \exp(-e^{\log \theta_{\text{cat}}[\omega] - m_\omega}) \cdot \exp\left(-\sum_{i \in W \setminus \omega} e^{\log \theta_{\text{cat}}[i] - m_\omega}\right) dm
\end{aligned}$$

Gumbel PDF.

$$\begin{aligned}
&= \int_{-\infty}^{\infty} \exp(\log \theta_{\text{cat}}[\omega] - m_\omega) \exp\left(-\sum_{i \in W} e^{\log \theta_{\text{cat}}[i] - m_\omega}\right) dm \\
&= \int_{-\infty}^{\infty} \theta_{\text{cat}}[\omega] \exp(-m_\omega) \exp\left(-e^{-m_\omega} \sum_{i \in W} \theta_{\text{cat}}[i]\right) dm \\
&= \pi_\omega \sum_{i \in W} \theta_{\text{cat}}[i] \int_{-\infty}^{\infty} \exp(-m_\omega) \exp\left(-e^{-m_\omega} \sum_{i \in W} \theta_{\text{cat}}[i]\right) dm \text{ From the above definition of } \pi_\omega \\
&= \pi_\omega \sum_{i \in W} \theta_{\text{cat}}[i] \frac{\exp\left(-e^{-m_\omega} \sum_{i \in W} \theta_{\text{cat}}[i]\right)}{\sum_{i \in W} \theta_{\text{cat}}[i]} \Big|_{-\infty}^{\infty} \\
&= \pi_\omega \sum_{i \in W} \theta_{\text{cat}}[i] \frac{1}{\sum_{i \in W} \theta_{\text{cat}}[i]} = \pi_\omega.
\end{aligned}$$

Reference: Huijben et al. (2022)

5 Future Directions

6 References

- Bai, Yunsheng, Hao Ding, Yang Qiao, Agustin Marinovic, Ken Gu, Ting Chen, Yizhou Sun, and Wei Wang. 2019. “Unsupervised Inductive Graph-Level Representation Learning via Graph-Graph Proximity.” <https://arxiv.org/abs/1904.01098>.
- Breiman, Leo. 2001. “Random Forests.” *Machine Learning* 45 (1): 5–32. <https://doi.org/10.1023/A:1010933404324>.
- Chen, Jialin, Shirley Wu, Abhijit Gupta, and Rex Ying. 2023. “D4Explainer: In-Distribution GNN Explanations via Discrete Denoising Diffusion,” no. arXiv:2310.19321 (October). <https://doi.org/10.48550/arXiv.2310.19321>.
- Debnath, Asim Kumar, Rosa L. Lopez de Compadre, Gargi Debnath, Alan J. Shusterman, and Corwin Hansch. 1991. “Structure-Activity Relationship of Mutagenic Aromatic and Heteroaromatic Nitro Compounds. Correlation with Molecular Orbital Energies and Hydrophobicity.” *Journal of Medicinal Chemistry* 34 (2): 786–97. <https://doi.org/10.1021/jm00106a046>.
- ERDdS, P, and A R&wi. 1959. “On Random Graphs i.” *Publ. Math. Debrecen* 6 (290-297): 18.
- Gallicchio, Claudio, and Alessio Micheli. 2019. “Fast and Deep Graph Neural Networks.” <https://arxiv.org/abs/1911.08941>.
- Gilbert, E. N. 1959. “Random Graphs.” *The Annals of Mathematical Statistics* 30 (4): 1141–44.
- Huijben, Iris A. M., Wouter Kool, Max B. Paulus, and Ruud J. G. van Sloun. 2022. “A Review of the Gumbel-Max Trick and Its Extensions for Discrete Stochasticity in Machine Learning,” no. arXiv:2110.01515 (March). <https://doi.org/10.48550/arXiv.2110.01515>.
- Maddison, Chris J., Andriy Mnih, and Yee Whye Teh. 2017. “The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables,” no. arXiv:1611.00712 (March). <https://doi.org/10.48550/arXiv.1611.00712>.
- Olah, Chris, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. 2020. “Zoom in: An Introduction to Circuits.” *Distill*. <https://doi.org/10.23915/distill.00024.001>.
- Rado, Richard. 1964. “Universal Graphs and Universal Functions.” *Acta Arithmetica* 9: 331–40. <https://api.semanticscholar.org/CorpusID:118279788>.
- Wang, Xiaoqi, and Han-Wei Shen. 2024. “GNNInterpreter: A Probabilistic Generative Model-

- Level Explanation for Graph Neural Networks,” no. arXiv:2209.07924 (February). <https://doi.org/10.48550/arXiv.2209.07924>.
- Xiao, Guanghua, Shidan Wang, Ruichen Rong, Donghan Yang, Xinyi Zhang, Xiaowei Zhan, Justin Bishop, et al. 2023. *Deep Learning of Cell Spatial Organizations Identifies Clinically Relevant Insights in Tissue Images*. Preprint. In Review. <https://doi.org/10.21203/rs.3.rs-2928838/v1>.
- Yuan, Hao, Haiyang Yu, Shurui Gui, and Shuiwang Ji. 2022. “Explainability in Graph Neural Networks: A Taxonomic Survey,” no. arXiv:2012.15445 (July). <https://doi.org/10.48550/arXiv.2012.15445>.
- Zhang, Zaixi, Qi Liu, Hao Wang, Chengqiang Lu, and Cheekong Lee. 2021. “ProtGNN: Towards Self-Explaining Graph Neural Networks,” no. arXiv:2112.00911 (December). <https://doi.org/10.48550/arXiv.2112.00911>.

7 Appendix