

# Learning and Evaluating Graph Neural Network Explanations based on Counterfactual and Factual Reasoning

Juntao Tan, Shijie Geng, Zuohui Fu, Yingqiang Ge, Shuyuan Xu, Yunqi Li, Yongfeng Zhang

Department of Computer Science, Rutgers University, New Brunswick, NJ 08854, US

{juntao.tan, shijie.geng, zuohui.fu, yingqiang.ge, shuyuan.xu, yunqi.li, yongfeng.zhang}@rutgers.edu

## ABSTRACT

Structural data well exists in Web applications, such as social networks in social media, citation networks in academic websites, and threads data in online forums. Due to the complex topology, it is difficult to process and make use of the rich information within such data. Graph Neural Networks (GNNs) have shown great advantages on learning representations for structural data. However, the non-transparency of the deep learning models makes it non-trivial to explain and interpret the predictions made by GNNs. Meanwhile, it is also a big challenge to evaluate the GNN explanations, since in many cases, the ground-truth explanations are unavailable.

In this paper, we take insights of Counterfactual and Factual ( $CF^2$ ) reasoning from causal inference theory, to solve both the learning and evaluation problems in explainable GNNs. For generating explanations, we propose a model-agnostic framework by formulating an optimization problem based on both of the two causal perspectives. This distinguishes  $CF^2$  from previous explainable GNNs that only consider one of them. Another contribution of the work is the evaluation of GNN explanations. For quantitatively evaluating the generated explanations without the requirement of ground-truth, we design metrics based on Counterfactual and Factual reasoning to evaluate the *necessity* and *sufficiency* of the explanations. Experiments show that no matter ground-truth explanations are available or not,  $CF^2$  generates better explanations than previous state-of-the-art methods on real-world datasets. Moreover, the statistic analysis justifies the correlation between the performance on ground-truth evaluation and our proposed metrics. Source code is available at [https://github.com/chrisjtan/gnn\\_cff](https://github.com/chrisjtan/gnn_cff).

## KEYWORDS

Explainable AI; Graph Neural Networks; Counterfactual Explanation; Machine Learning; Machine Reasoning; Causal Inference

## ACM Reference Format:

Juntao Tan, Shijie Geng, Zuohui Fu, Yingqiang Ge, Shuyuan Xu, Yunqi Li, Yongfeng Zhang. 2022. Learning and Evaluating Graph Neural Network Explanations based on Counterfactual and Factual Reasoning. In *Proceedings of the ACM Web Conference 2022 (WWW '22)*, April 25–29, 2022, Virtual Event, Lyon, France. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3485447.3511948>

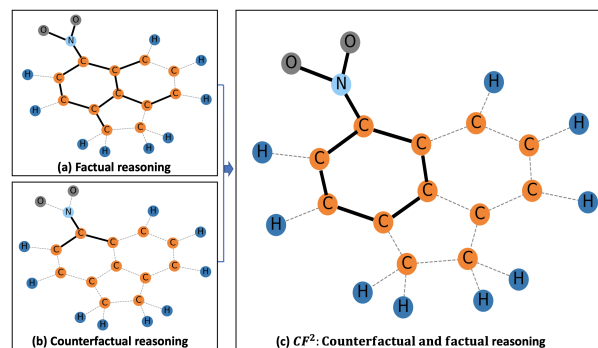
Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

WWW '22, April 25–29, 2022, Virtual Event, Lyon, France

© 2022 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-9096-5/22/04...\$15.00

<https://doi.org/10.1145/3485447.3511948>



**Figure 1: An example for extracting explanations for mutagenic prediction. The sub-graph induced by the bold edges is the explanation extracted by (a) factual reasoning, (b) counterfactual reasoning and (c) counterfactual and factual reasoning. The sub-graph in (c) is also the ground-truth explanation, i.e., Nitrobenzene structure is the cause of mutagen.**

## 1 INTRODUCTION

Structured data widely exists in various domains such as social networks [42], citation networks [14, 33] in Web applications, and chemical molecules [9, 40] in biomedical research. Such kind of data, which is commonly represented as graph, contains rich information. However, conducting studies on graph data is exhausting for human because both the topology information and the node features need to be considered.

Fortunately, GNNs have shown great advantages on learning graph representations because they aggregate both the feature and structure information by passing the messages in the graph. Thus, GNN-based models achieved promising results in graph prediction tasks such as graph classification, node classification, and link prediction. However, most of the GNN models are non-transparent, which leads to the lack of explainability in model predictions. Exploring the explainability of GNNs is crucial because good explanations not only help to understand the model predictions but also help to identify potential flaws in the model and further refine the GNN model.

In a high-level view, recent state-of-the-art GNN explanation methods are based on either factual reasoning [24, 43] or counterfactual reasoning [22, 23]. Methods based on factual reasoning seek a sub-graph whose information is *sufficient* to produce the same prediction as using the whole original graph, while methods based on counterfactual reasoning seek a sub-graph whose information is *necessary* which if removed will result in different predictions.

Both factual reasoning and counterfactual reasoning are important approaches to explanation extraction, but each of them alone

has its disadvantages. Factual reasoning favors sub-graph explanations that contain enough information to make the same prediction, but the extracted sub-graph may include redundant nodes/edges and thus not compact enough. For example, an extreme case is to take the whole graph as the “sub-graph,” which will definitely give the same prediction, but such a “sub-graph” does not convey any meaningful information as an explanation.

This disadvantage is also illustrated in Figure 1(a). We use a real-world biochemical example since it has known ground-truth explanation which is hardly accessible for most Web-based graphs. In this example, a molecule is predicted to be mutagenic and we want to extract explanations for the prediction. The explanation sub-graph generated by factual reasoning may indeed cover the essential reason—the Nitrobenzene structure (benzene-NO<sub>2</sub>) [9]. However, it also contains some extra edges from other carbon rings because when these edges are included, the sub-graph leads to the same mutagenic prediction. In a nutshell, the extracted explanation tends to be sufficient but not necessary.

On the other hand, counterfactual reasoning favors the explanations that only contain the most crucial information, i.e., if the explanation sub-graph is removed, then the graph will result in different predictions. However, because of the this, counterfactual reasoning may only extract a small subset of the real explanation.

Take Figure 1(b) as an example, counterfactual reasoning generates a sub-graph with only three edges. These edges, if removed, will indeed break the Nitrobenzene structure and thus lead to a different prediction (i.e., non-mutagenic), however, such an explanation does not cover the complete information about what makes the target molecule mutagenic. In a nutshell, the extracted explanation tends to be necessary but not sufficient.

To overcome the problems and to seek a balance between necessity and sufficiency, we propose a Counterfactual and Factual (CF<sup>2</sup>) reasoning framework to extract GNN explanations which brings the best of the two worlds. CF<sup>2</sup> formulates an optimization problem to integrate counterfactual and factual reasoning objectives so as to extract explanations that are both necessary and sufficient. As shown in Figure 1(c), the counterfactual objective encourages the necessary edges while the factual objective ensures that the extracted explanation contains sufficient information, and thus an ideal sub-graph explanation can be induced.

Another challenge in explainable GNN research is that most real-world graph datasets lack ground-truth explanations, which makes it difficult to evaluate the extracted explanations for these datasets. Fortunately, the fundamental idea of CF<sup>2</sup> can also be adapted into the evaluations. In this paper, we borrow insights from causal inference theory and adopt the *Probability of Necessity* (PN) and *Probability of Sufficiency* (PS) to evaluate the necessity and sufficiency of the extracted explanations, which makes it possible to conduct quantitative evaluation of GNN explanations. PN and PS are aligned with counterfactual and factual reasoning respectively. Details are formulated in Section 6.

In summary, this work has the following contributions:

- We show the relationship between factual (or counterfactual) reasoning and the sufficiency (or necessity) of GNN explanations.
- We propose a CF<sup>2</sup> framework to consider both factual and counterfactual reasoning for GNN explanations.

- We propose a set of quantitative evaluation metrics to evaluate the GNN explanations.
- We conduct extensive experiments on 2 synthetic datasets and 3 real-world datasets from different domains to justify the proposed model and evaluation metric.

## 2 RELATED WORKS

### 2.1 Explainability in Deep Learning and AI

Explainable AI has been an important topic in recommender systems [5, 6, 13, 36, 41, 46, 47], natural language processing [8, 16, 20] and computer vision [7, 10, 15, 25, 38]. To improve the transparency of deep neural networks, many explanation techniques have been proposed in recent years. Based on how to obtain the importance scores, these approaches can be categorized into gradient/feature-based methods, perturbation/casual-based methods, and surrogate methods [26, 45]. Gradients/feature-based methods [19, 32, 35] are the most straightforward way to achieve saliency maps as explanations. They usually map the final prediction to the input space by gradient back-propagation or by linking hidden features to inputs via interpolation. Perturbation/casual-based methods [11, 12, 15, 27, 36, 37, 39] learn the feature importance through observing the change of predictions with respect to the input perturbation. The idea behind these methods are intuitive: determining which part of the inputs are important by either removing the least important information (i.e., pixels in image, words in text, nodes in graph) to keep the model prediction the same (factual reasoning) or removing the most important information to change the model prediction (counterfactual reasoning). The representative of surrogate methods is LIME [31], which employs a simple linear model to approximate the predictions on a bunch of nearby inputs and provides explanations from the surrogate model.

### 2.2 Explainability in Graph Neural Networks

The aforementioned methods are developed mainly for images and texts. Besides the individual features, graphs also contain important topological structure. Such graph structures are highly related to the functionalities in specific domains and should not be ignored for GNN-based explanation approaches. In explainable GNN, early attempts directly extend gradients/feature-based methods [1, 29] to identify important input features. While simple and efficient, these approaches either suffer from gradient saturation [34] or lack of the ability to explain node classification predictions [45]. Another line of work [17] follows LIME and adopts a surrogate model for explaining deep graph models. But it ignores the graph structure and cannot explain graph classification models. Hence, these approaches are not suitable for explaining the graph-level predictions of GNNs. To solve the problem, Ying et al. [43] proposed GNNExplainer which treats explanation generation as a mask optimization problem. It follows the idea in perturbation/casual-based methods and learns soft masks that cover the key nodes and edges while maintaining the original prediction score. GISST [21] further extended GNNExplainer by identifying important sub-graphs and generating importance scores for all nodes and edges through a self-attention layer. The above two methods learn soft masks that contain continuous values, which suffer from the “introduce evidence” problem [45]. To solve the problem, PGExplainer [24] adopts

the reparameterization trick and learns approximate discrete masks that maximizes the mutual information between key structures and predictions, and XGNN [44] generates a graph based on reinforcement learning to approximate the prediction of the original graph. As generative models, they also facilitate the holistic explanation for multiple instances. Apart from these factual reasoning approaches, there are also recent works exploring counterfactual reasoning. CF-GNNExplainer [23] introduces counterfactual reasoning to renovate GNNExplainer and is able to generate minimal yet crucial explanations for GNNs. Gem [22] distills ground-truth explanations based on Granger causality (a type of counterfactual reasoning) and then trains an auto-encoder architecture to generate adjacency matrix as explanations based on supervised learning. However, these GNN-based explanation approaches only consider factual or counterfactual reasoning alone, and thus will bias towards either sufficiency or necessity rather than achieving a balance when extracting explanations. In this paper, we seek to integrate counterfactual and factual reasoning to extract GNN explanations that are both sufficient and necessary.

### 3 PRELIMINARIES AND NOTATIONS

In this section, we briefly introduce how GNNs learn the node and graph representations, as well as its application in the node classification and graph classification tasks. We also introduce the basic notations to be used throughout the paper.

#### 3.1 Learning Representations

Given a graph  $G = \{\mathcal{V}, \mathcal{E}\}$ , and each node  $v_i \in \mathcal{V}$  has a  $d$ -dimensional node feature  $x_i \in \mathbb{R}^d$ . GNN learns the representation of  $v_i$  by iteratively aggregating the information of its neighbors  $N(i)$ . At the  $l$ -th layer of a GNN model,  $v_i$ 's representation  $h_i = \text{update}(h_i^{l-1}, h_{N(i)}^l)$ , where  $h_i^{l-1}$  is the representation of  $v_i$  in the previous layer, and  $h_{N(i)}^l$  is aggregated from the neighbors of  $v_i$  via an aggregation function:  $h_{N(i)}^l = \text{aggregate}(h_j^{l-1}, \forall v_j \in N(i))$ . The implantation of the  $\text{update}(\cdot)$  and  $\text{aggregate}(\cdot)$  functions can be different for different GNN models. For a GNN model with  $L$  layers in total,  $h_i^L$  is the final representation of the node  $v_i$ .

After aggregating the node representations, the graph representation can be computed by taking the average of all the node representations in the graph.

#### 3.2 Graph Classification

Given a set of  $n$  graphs  $\mathcal{G} = \{G_1, G_2, \dots, G_n\}$ , and each graph  $G_k \in \mathcal{G}$  is associated with a ground-truth class label  $y_k \in C$ , where  $C = \{1, 2, \dots, r\}$  is the set of graph classes. The graph classification task aims to learn a graph classifier  $\Phi$  that predicts the estimated label  $\hat{y}_k$  for an input graph  $G_k$ .

Each input graph  $G_k = \{\mathcal{V}_k, \mathcal{E}_k\}$  is associated with an adjacency matrix  $A_k \in \{0, 1\}^{|\mathcal{V}_k| \times |\mathcal{V}_k|}$  and a node feature matrix  $X_k \in \mathbb{R}^{|\mathcal{V}_k| \times d}$ . After the training process, the GNN model will predict the estimated label  $\hat{y}_k$  for  $G_k$  by:

$$\hat{y}_k = \arg \max_{c \in C} P_\Phi(c | A_k, X_k) \quad (1)$$

where  $\Phi$  is the trained GNN model.

### 3.3 Node Classification

For the node classification task, the goal is to predict the class label for each node in a given graph  $G = \{\mathcal{V}, \mathcal{E}\}$ . Each node  $v_i \in \mathcal{V}$  is associated with a ground-truth node label  $y_i \in C$ , where  $C = \{1, 2, \dots, r\}$  is the set of node classes. In node classification task, since only the  $L$ -hop neighbors of the node  $v_i$  will influence  $h_i^L$ , we define the  $L$ -hop sub-graph of the node  $v_i$  as  $G_{s(i)}$  which is the computational graph that will be the input of the GNN model.  $A_{s(i)}$  and  $X_{s(i)}$  are the related adjacency matrix and feature matrix of the computational sub-graph. The trained GNN model will thus predict the estimated label  $\hat{y}_i$  for the node  $v_i$  as:

$$\hat{y}_i = \arg \max_{c \in C} P_\Phi(c | A_{s(i)}, X_{s(i)}) \quad (2)$$

## 4 PROBLEM FORMULATION

In this section, we first introduce the explainable GNN problem for the classification task. Then, we mathematically define two objectives for extracting explanations and adjust them into the CF<sup>2</sup> framework. The two objectives are 1) an effective explanation should be both *sufficient* and *necessary*, which are reflected by the factual and counterfactual conditions, respectively; and 2) a good explanation should not only be *effective*, but also be *simple*, which is driven by the Occam's Razor Principle [2]. We formulate the Explanation Strength to reflect the effectiveness and formulate the Explanation Complexity to reflect the simpleness. The above two objectives are the foundation of the CF<sup>2</sup> framework for extracting explanations.

We note that in the rest of the paper, all the concepts, examples and mathematical definitions are introduced under the **graph classification** problem setting and they can be easily generalized to the node classification task. We provide another version for node classification in Appendix A.

#### 4.1 Explainable Graph Neural Networks

Suppose a graph  $G_k = \{\mathcal{V}_k, \mathcal{E}_k\}$  has the predicted label  $\hat{y}_k$ , following the setup of Ying et al. [43], we generate the explanation for this prediction as a sub-graph, which consists of a subset of the edges and a subset of the feature space of the original graph. The sub-graph can be either connected or unconnected. Thus, the goal of CF<sup>2</sup> is to learn an edge mask  $M_k \in \{0, 1\}^{|\mathcal{V}_k| \times |\mathcal{V}_k|}$  and a feature mask  $F_k \in \{0, 1\}^{|\mathcal{V}_k| \times d}$ , which will be applied on the adjacency matrix  $A_k \in \{0, 1\}^{|\mathcal{V}_k| \times |\mathcal{V}_k|}$  and the node feature matrix  $X_k \in \mathbb{R}^{|\mathcal{V}_k| \times d}$  of the original graph  $G_k$ . After optimization, the sub-graph will be  $A_k \odot M_k$  with the sub-features  $X_k \odot F_k$ , which is the generated explanation for the prediction of graph  $G_k$ .

#### 4.2 Counterfactual and Factual Conditions

As discussed above, an ideal explanation should be both necessary and sufficient. CF<sup>2</sup> achieves this goal by considering both factual and counterfactual reasoning.

Factual and counterfactual reasoning are two opposite but very symmetric ways of reasoning. Factual reasoning asks the question "Given A already happened, will B happen?" Counterfactual reasoning, on the contrary, asks "If A did not happen, will B still happen?" [30]. Under the context of GNN explanations, factual reasoning generates sub-edges/sub-features that satisfy the condition "With

these sub-edges/sub-features, which is consistent with the fact, the GNN prediction will be the same.” Counterfactual reasoning generates sub-edges/sub-features that satisfy the condition “Without these sub-edges/sub-features, which is inconsistent with the fact, the GNN prediction will be different.” Intuitively, factual reasoning seeks a *sufficient* set of edges/features that produce the same prediction as using the whole graph, while counterfactual reasoning seeks a *necessary* set of edges/features that if removed will lead to different predictions.

In  $CF^2$ , both factual and counterfactual reasoning are formulated into the model. The condition for factual reasoning is mathematically formulated as following:

$$\begin{aligned} &\text{Condition for Factual Reasoning :} \\ &\arg \max_{c \in C} P_{\Phi}(c \mid A_k \odot M_k, X_k \odot F_k) = \hat{y}_k \end{aligned} \quad (3)$$

Similarly, the condition for counterfactual reasoning is formulated as:

$$\begin{aligned} &\text{Condition for Counterfactual Reasoning :} \\ &\arg \max_{c \in C} P_{\Phi}(c \mid A_k - A_k \odot M_k, X_k - X_k \odot F_k) \neq \hat{y}_k \end{aligned} \quad (4)$$

These two conditions will be reflected as objectives for explanation extraction in the loss function, which will be introduced in Section 5.

### 4.3 Simple and Effective Explanations

According to the Occam’s Razor Principle [2], if two explanations are equally effective, we tend to prefer the simpler one. To achieve this goal, we introduce Explanation Complexity and Explanation Strength for GNN explanations. These two concepts help  $CF^2$  to seek simple and effective explanations for GNN predictions.

Explanation complexity  $C(M, F)$  measures how complicated the explanation is, which is defined as the number of edges/features used to construct the explanation. Note that  $M$  and  $F$  are binary matrices indicating which edges and features are included in the sub-graph explanation. As a result,  $C(M, F)$  can be defined as the number of 1’s in  $M$  and  $F$  matrices, i.e.,

$$C(M, F) = \|M\|_0 + \|F\|_0 \quad (5)$$

However, to make  $C(M, F)$  optimizable, we will relax it from 0-norm to 1-norm. We will explain in Section 5.

Explanation strength  $S(M, F)$  measures how effective the explanation is. As mentioned above, an effective explanation should be both sufficient and necessary, which is pursued by the factual and counterfactual conditions (Eq.(3) and (4)). As a result, the explanation strength can be defined as two parts: factual explanation strength  $S_f(M, F)$  and counterfactual explanation strength  $S_c(M, F)$ , both are the larger the better.

The mathematical definition of  $S_f(M, F)$  is consistent with the condition for factual reasoning, which is:

$$S_f(M, F) = P_{\Phi}(\hat{y}_k \mid A_k \odot M_k, X_k \odot F_k) \quad (6)$$

On the contrary,  $S_c(M, F)$  is consistent with the condition for counterfactual reasoning, which is:

$$S_c(M, F) = -P_{\Phi}(\hat{y}_k \mid A_k - A_k \odot M_k, X_k - X_k \odot F_k) \quad (7)$$

**Table 1:  $CF^2$  generates explanations with two goals: 1) the explanation should be simple, i.e., low in explanation complexity, which means that the generated explanation sub-graph should have a small number of edges and features, which can be achieved by 0-norm or 1-norm regularization. 2) the explanation should be effective, i.e., high in explanation strength. An effective explanation should be both sufficient and necessary. Sufficiency can be achieved via factual reasoning and necessity via counterfactual reasoning.**

Objs	Simple (↓ Complexity)	Effective (↑ Strength)	
Measure	# edges, # features	Sufficiency	Necessity
Method	Regularization	Factual	Counterfactual

Explanation complexity and strength will serve as the learning objective and learning constraint in the explanation extraction algorithm, which will also be introduced in Section 5.

In Table 1, we provide an overview of the relationships among the aforementioned concepts.

## 5 THE $CF^2$ FRAMEWORK

In this section, we first introduce the  $CF^2$  constrained optimization framework. Then we provide a relaxed version to make the framework optimizable.

### 5.1 $CF^2$ Optimization Problem

$CF^2$  is able to generate explanation for any prediction made by a GNN model. As mentioned before,  $CF^2$  aims to find simple (i.e., low complexity) and effective (i.e., high strength) explanations, which can be shown as the following constrained optimization framework:

$$\begin{aligned} &\text{minimize Explanation Complexity} \\ &\text{s.t., Explanation is Strong Enough} \end{aligned} \quad (8)$$

According to the mathematical definition of explanation complexity and strength in Section 4.3, for a given graph  $G_k$  with predicted label  $\hat{y}_k$ , Eq.(8) can be rewritten as:

$$\begin{aligned} &\text{minimize } C(M_k, F_k) \\ &\text{s.t., } S_f(M_k, F_k) > P_{\Phi}(\hat{y}_{k,s} \mid A_k \odot M_k, X_k \odot F_k), \\ &S_c(M_k, F_k) > -P_{\Phi}(\hat{y}_{k,s} \mid A_k - A_k \odot M_k, X_k - X_k \odot F_k) \end{aligned} \quad (9)$$

where  $\hat{y}_{k,s}$  is the label other than  $\hat{y}_k$  that has the largest probability score predicted by the GNN model. Intuitively, the constraint aims to ensure that when only using the information in the explanation sub-graph, the predicted label  $\hat{y}_k$ ’s probability is higher than any other label and thus the prediction does not change, while if information in the explanation sub-graph is removed,  $\hat{y}_k$ ’s probability will be smaller than at least one other label and thus the prediction will change.

### 5.2 Relaxed Optimization

Directly optimizing Eq.(9) is challenging because both the objective part and the constraint part are not differentiable. As a result, we relax the two parts to make them optimizable.

For the objective part, we relax the masks  $M_k$  and  $F_k$  to real values, which are  $M_k^* \in \mathbb{R}^{|\mathcal{V}_k| \times |\mathcal{V}_k|}$  and  $F_k^* \in \mathbb{R}^{|\mathcal{V}_k| \times d}$ . Meanwhile, since the 0-norm in the original equation is also not differentiable,

we use 1-norm to ensure the sparsity of  $M_k^*$  and  $F_k^*$ , which has been proven to be effective in [3, 4].

For the constraint part, we relax it as pairwise contrastive loss  $L_f$  and  $L_c$ , where

$$L_f = \text{ReLU}(\gamma + P_\Phi(\hat{y}_{k,s} \mid A_k \odot M_k^*, X_k \odot F_k^*) - S_f(M_k^*, F_k^*)) \quad (10)$$

Similarly,

$$L_c = \text{ReLU}(\gamma - S_c(M_k^*, F_k^*) - P_\Phi(\hat{y}_{k,s} \mid A_k - A_k \odot M_k^*, X_k - X_k \odot F_k^*)) \quad (11)$$

After relaxation, Eq.(9) becomes optimizable, which is:

$$\text{minimize } \|M_k^*\|_1 + \|F_k^*\|_1 + \lambda(\alpha L_f + (1 - \alpha)L_c) \quad (12)$$

When solving the relaxed optimization equation, the margin value  $\gamma$  in Eq.(10) and Eq.(11) is set to 0.5. After the optimization, 0.5 is also used as the threshold to be applied on the optimized masks to generate explanations (i.e., when the value in the masks  $M^*/F^*$  is larger than 0.5, we keep the related edge/feature in the generated explanation).

In Eq.(12), the hyper-parameter  $\lambda$  controls the trade-off between the explanation complexity and the explanation strength. By increasing  $\lambda$ , the model will focus more on the effectiveness of the generated explanations but less on the complexity, which may result in a bigger sub-graph and feature space. Another hyper-parameter  $\alpha$  controls the trade-off between the sufficiency and the necessity of the generated explanation. By increasing (or decreasing)  $\alpha$ , the generated explanation will focus more on the sufficiency (or necessity).

## 6 EVALUATING GNN EXPLANATIONS

Most of the real-world datasets for graph/node classification do not have ground-truth explanations, which makes the evaluation of GNN explanations a big challenge for the community. As mentioned in section 4, a good explanation should be both sufficient and necessary, which is aligned with the factual and counterfactual condition, respectively.

In logic and mathematics, necessity and sufficiency are terms used to describe a conditional or implicational relationship between two statements. Suppose we have  $S \Rightarrow N$ , i.e., if  $S$  happens then  $N$  will happen, then we say  $S$  is a sufficient condition for  $N$ . Meanwhile, we have the logically equivalent contrapositive  $\neg N \Rightarrow \neg S$ , i.e., if  $N$  does not happen, then  $S$  will not happen, as a result, we say  $N$  is a necessary condition for  $S$ . In light of this idea, we adopt the concepts of Probability of Sufficiency (PS) and Probability of Necessity (PN) from causal inference theory [28, p.112], which enable us to conduct quantitative evaluation of the GNN explanations.

### 6.1 Probability of Sufficiency

For an explanation  $A$  that is generated to explain event  $B$ , suppose  $A$  happens then  $B$  will happen, then  $A$  satisfies the factual condition and  $A$  is a sufficient explanation. We define PS as the percentage of generated explanations that are *sufficient* for the instance to achieve the same prediction as using the whole graph. In explainable GNN

problem, **Probability of Sufficiency** is defined as:

$$\text{PS} = \frac{\sum_{G_k \in \mathcal{G}} \text{ps}_k}{|\mathcal{G}|}, \text{ where } \text{ps}_k = \begin{cases} 1, & \text{if } \hat{y}'_k = \hat{y}_k \\ 0, & \text{else} \end{cases} \quad (13)$$

where  $\hat{y}'_k = \arg \max_{c \in C} P_\Phi(c \mid A_k \odot M_k, X_k \odot F_k)$

Intuitively, PS measures the percentage of graphs whose explanation sub-graph alone can keep the GNN prediction unchanged, and thus it is sufficient.

### 6.2 Probability of Necessity

Similarly, suppose  $A$  does not happen then  $B$  will not happen, we say  $A$  satisfies the counterfactual condition and  $A$  is a necessary explanation. We define PN as the percentage of generated explanations that are *necessary* for the instance to achieve the same prediction as using the whole graph. In explainable GNN problem, **Probability of Necessity** is defined as:

$$\text{PN} = \frac{\sum_{G_k \in \mathcal{G}} \text{pn}_k}{|\mathcal{G}|}, \text{ where } \text{pn}_k = \begin{cases} 1, & \text{if } \hat{y}'_k \neq \hat{y}_k \\ 0, & \text{else} \end{cases} \quad (14)$$

where  $\hat{y}'_k = \arg \max_{c \in C} P_\Phi(c \mid A_k - A_k \odot M_k, X_k - X_k \odot F_k)$

Intuitively, PN measures the percentage of graphs whose explanation sub-graph, if removed, will change the GNN prediction, and thus it is necessary.

Both PS and PN are the higher the better. Similar to the definition of  $F_1$  score, we use  $F_{NS} = \frac{2 \cdot \text{PN} \cdot \text{PS}}{\text{PN} + \text{PS}}$  to measure the overall performance of a GNN explanation method.

## 7 EXPERIMENTS

In this section, we first introduce the datasets and the comparison baselines. Then, we report the main experimental results and the analyses. Finally, we conduct experiments to show the influence of factual and counterfactual reasoning, which helps to gain deeper understanding of the key concepts of the paper. We also conduct studies to justify the effectiveness of the PN/PS-based evaluation.

### 7.1 Datasets

We test our algorithm on two synthetic and three real-world datasets. The two synthetic datasets are BA-shapes and Tree-Cycles, which were introduced in Ying et al. [43]. We follow exactly the same setup when generating these two datasets. The three real-world datasets are Mutag [9], NCI1 [40] and CiteSeer [14, 33]. The Mutag dataset contains 4,337 molecules classified into two categories: mutagenic or non-mutagenic. The NCI1 dataset contains 4,110 chemical compounds which are categorized as either positive or negative to cell lung cancer. The CiteSeer dataset contains 3,312 scientific publications classified into six classes, in which the nodes are the papers and the links represent that one paper is cited by another one.

BA-Shapes, Tree-Cycles and CiteSeer are for node classification, while Mutag and NCI1 are for graph classification. BA-Shapes and Tree-Cycles have ground-truth motifs (i.e., ‘‘house’’ and ‘‘cycle’’ structures) for explaining the classification since they are human-designed. However, NCI1 and CiteSeer do not have such ground-truth motifs. We would like to especially mention the motifs in the Mutag dataset. Luo et al. [24] assumed that the nitro group

(NO<sub>2</sub>) and amino group (NH<sub>2</sub>) are the true reasons for mutagenicity and filtered out the mutagens that do not contain them. However, according to Debnath et al. [9], which is the work that published the Mutag dataset, NH<sub>2</sub> requires microsomal activation to achieve full mutagenic potency and the dataset is limited to studies without such activation. Thus, NH<sub>2</sub> has very small influence in the Mutag dataset. This is also mentioned in Lin et al. [22], which shows that the presence of NH<sub>2</sub> has very low correlation with the classification result on this dataset. In fact, benzene-NO<sub>2</sub> is the only discriminative motif in this dataset. As a result, we extract a sub-dataset, Mutag<sub>0</sub>, which only includes those chemical compounds that contain benzene-NO<sub>2</sub> and are mutagenic, or that does not contain benzene-NO<sub>2</sub> and are not mutagenic. The statistics of the Mutag dataset are shown in Table 2. Table 3 provides the statistics of all the datasets used.

## 7.2 Baselines

The comparable baselines in this paper should satisfy such conditions: 1) They generate sub-graphs for explanation; 2) They can generate explanations for any graph dataset, with or without prior knowledge, e.g., Luo et al. [24] requires explicit motif to generate explanations thus could not be applied on NCI1 and CiteSeer, which is the reason why it is not included. The baselines are as follows:

**GNNEExplainer** [43]: An explanation model base on perturbation. It selects a compact sub-graph while maximizing the mutual information with the whole graph.

**CF-GNNEExplainer** [23]: An extension of GNNEExplainer by generating explanations based on counterfactual reasoning.

**Gem** [22]: A generative explanation model based on Granger causality, it trains auto-encoder to generate explanation sub-graphs.

## 7.3 Experimental Setup

There are two phases in the experiments: 1) Training the base GNN model for classification; and 2) Generating the explanations.

For the base model, a GCN with three layers is used for all the datasets. The hidden dimensions are 16 for BA-Shapes, Tree-Cycles, Mutag and NCI1, and 32 for CiteSeer. The model for Mutag and NCI1 datasets requires an extra pooling and fully convolution layers for computing the graph embeddings. We apply ReLU activation function after all the layers except for the last layer, which is followed by a Softmax function for classification. The learning rate is 0.001 during training for all datasets and the ratio between training and test set is 8 : 2. In Table 6, we report the number of training epochs and the accuracy of the base model we used in this paper. We use the same base model for all the baselines to fairly compare the explanation ability. Since the explanation method is model-agnostic, the base model can be any classification model for graphs.

In the explanation phase, GNNEExplainer and Gem require a human-selected  $K$  value to decide the size of the explanations in their settings. When implementing these two methods, we follow the same setup in Gem: for the synthetic datasets, we set  $K$  equal to the size (#edges) of the ground-truth motifs, and we set  $K = 15$  (#edges) for Mutag and NCI1. We run two experiments on the CiteSeer dataset: edge-based explanation ( $K = 5$ ) and feature-based explanation ( $K = 60$ ). CF-GNNEExplainer and CF<sup>2</sup> do not require

**Table 2: Statistics of the Mutag dataset, the molecules with “\*” are the graphs we used to build the Mutag<sub>0</sub> dataset.**

	w/ benzene-NO <sub>2</sub>	w/o benzene-NO <sub>2</sub>
mutagen	448*	1,953
non-mutagen	83	1,853*

**Table 3: Statistics of all datasets. “#ave n” and “#ave e” are the number of nodes/edges per graph. “#feat” is the number of features. In the “task” column, “node” and “graph” indicate the dataset is used for the node classification task or graph classification task, respectively. The check marks in the “gt” column means the existence of ground-truth motifs.**

Dataset	#graph	#ave n	#ave e	#class	#feat	task	gt
BA-Shapes	1	700	4100	4	-	node	✓
Tree-Cycles	1	871	1950	2	-	node	✓
Mutag	4337	30.32	30.77	2	14	graph	
Mutag <sub>0</sub>	2301	31.74	32.54	2	14	graph	✓
NCI1	4110	29.87	32.30	2	37	graph	
CiteSeer	1	3312	4732	6	3703	node	

prior knowledge about the  $K$  value. The size of explanations are automatically decided by the model themselves via optimization.

For the hyper-parameters in CF<sup>2</sup>, the  $\lambda$  is decided by normalizing the 1-norm loss and the pairwise contrastive loss into the same scale, which are [500, 500, 1000, 20, 100] for BA-Shapes, Tree-Cycles, Mutag<sub>0</sub>, NCI1, and CiteSeer, respectively. For the  $\alpha$  value, we set it to be 0.6 to make factual reasoning slightly leading the optimization. We will conduct ablation study on  $\alpha$  in Section 7.6 to show its influence.

We evaluate the explanation methods based on the graphs in the test dataset. Since the BA-Shapes, Tree-Cycles and Mutag<sub>0</sub> datasets have ground-truth explanations, we report the Accuracy, Precision, Recall and  $F_1$  scores of the generated explanations of each method. Besides, for all datasets, we evaluate the explanation model with the PS, PN and  $F_{NS}$  metrics introduced in Section 6. Note that we not only generate explanations based on the edges, but also generate explanations on the node features and test them on the CiteSeer dataset, which is not examined in previous works.

## 7.4 Quantitative Analysis

In Table 4, we report the evaluation of the generated explanations with respect to the ground-truth motifs. CF<sup>2</sup> has an overall better performance than all the other baselines according to Accuracy and  $F_1$  scores. The only exception is when comparing with Gem on the BA-Shapes dataset with respect to Accuracy, which is lower by 0.62%. However, since Gem requires the size of the ground-truth motif to select exactly the same size of explanation, which is a strong prior knowledge, this minor difference is considered acceptable. Another observation is that CF-GNNEExplainer is higher in Precision and GNNEExplainer is higher in Recall when comparing with each other. This justifies our initial motivation about factual and counterfactual reasoning: The factual reasoning focuses on the sufficiency of the explanation, which results in a higher coverage on the ground-truth motifs, while counterfactual reasoning focuses on the necessity, which provides more precise explanations but

**Table 4: Explanation evaluation w.r.t ground-truth. Acc, Pr and Re represent Accuracy, Precision and Recall, respectively. Models with  $\dagger$  are the models that fix the size of explanations with pre-defined  $K$  values. For the metrics that measure the overall explanation performance (e.g.,  $F_1$  score), we use bold font to mark the highest scores. For the metrics that only measure partial performance (e.g., precision, recall), we mark the highest scores with underlines.**

Models	BA-Shapes				Tree-Cycles				Mutag <sub>0</sub>			
	Acc%	Pr%	Re%	$F_1$ %	Acc%	Pr%	Re%	$F_1$ %	Acc%	Pr%	Re%	$F_1$ %
GNNExplainer $\dagger$	95.25	60.08	60.08	60.08	92.78	68.06	68.06	68.06	96.96	59.71	85.17	68.85
CF-GNNExplainer	94.39	67.19	54.11	56.79	90.27	<u>87.40</u>	47.45	59.10	96.91	<u>66.09</u>	39.46	47.39
Gem $\dagger$	<b>96.97</b>	64.16	64.16	64.16	89.88	57.23	57.23	57.23	96.43	63.12	47.11	54.68
CF <sup>2</sup>	96.37	<u>73.15</u>	<u>68.18</u>	<b>66.61</b>	<b>93.26</b>	84.92	<u>73.84</u>	<b>75.69</b>	<b>97.34</b>	65.28	<u>88.59</u>	<b>72.56</b>

**Table 5: Explanation evaluation on PN/PS-based metrics. #exp is the size of the generated explanations. Models with  $\dagger$  are the models that fix the size of explanations with pre-defined  $K$  values. For the metrics that measure the overall explanation performance (e.g.,  $F_{NS}$  score), we use bold font to mark the highest scores. For the metrics that only measure partial performance (e.g., PN, PS), we mark the highest scores with underlines.**

Models	BA-Shapes				Tree-Cycles				Mutag <sub>0</sub>			
	PN%	PS%	$F_{NS}$ %	#exp	PN%	PS%	$F_{NS}$ %	#exp	PN%	PS%	$F_{NS}$ %	#exp
GNNExplainer $\dagger$	72.19	45.62	55.91	6.00	100.00	59.72	74.78	6.00	71.79	97.44	82.67	15.00
CF-GNNExplainer	75.34	41.10	53.18	5.79	100.00	31.94	48.42	3.44	96.26	7.48	13.88	7.72
Gem $\dagger$	61.36	52.27	56.45	6.00	100.00	29.89	46.02	6.00	83.01	76.42	79.58	15.00
CF <sup>2</sup>	<u>76.73</u>	<u>68.22</u>	<b>72.07</b>	6.21	<u>100.00</u>	<u>81.94</u>	<b>90.08</b>	5.81	<u>97.44</u>	<u>100.00</u>	<b>98.70</b>	14.95

Models	NCI1				CiteSeer (edge)				CiteSeer (feature)			
	PN%	PS%	$F_{NS}$ %	#exp	PN%	PS%	$F_{NS}$ %	#exp	PN%	PS%	$F_{NS}$ %	#exp
GNNExplainer $\dagger$	92.13	62.16	74.24	15.00	66.67	90.05	76.61	5.00	71.64	<u>99.50</u>	72.79	60.00
CF-GNNExplainer	97.14	31.43	47.49	7.75	69.50	82.00	75.23	2.58	72.14	92.54	81.07	72.91
Gem $\dagger$	99.03	52.15	68.32	15.00	61.05	72.67	66.36	5.00	-	-	-	-
CF <sup>2</sup>	<u>100.00</u>	<u>63.81</u>	<b>77.91</b>	17.70	<u>71.00</u>	<u>94.50</u>	<b>81.08</b>	3.18	<u>74.63</u>	95.02	<b>83.60</b>	62.73

worse in coverage. As a result, CF<sup>2</sup> is balancing between them and has an overall higher performance in  $F_1$ .

Then, for all the datasets, we test the generated explanations with the PN, PS, and  $F_{NS}$  scores, as shown in Table 5. CF<sup>2</sup> performs the best among all the baselines on PN in 100% cases, on PS in 83% cases, and on  $F_{NS}$  in 100% cases. Moreover, CF<sup>2</sup> has 13.57% average improvement than the best performance of the baselines on  $F_{NS}$ , which is significant. Similar to the observations in the ground-truth evaluation, we note that the counterfactual-based methods perform better in PN and factual-based methods perform better in PS. This is in line with our previous analysis on the advantages and disadvantages of factual and counterfactual reasoning. Besides, this result also gives us insights about the relationship between Precision/Recall and PN/PS.

## 7.5 Qualitative Analysis

In Figure 2, we illustrate explanations based on topology structures to qualitatively compare CF<sup>2</sup> with the methods based on only factual (GNNExplainer) or counterfactual (CF-GNNExplainer) reasoning. Results show that CF<sup>2</sup> better discovers graph motifs than the other two methods. Moreover, counterfactual-based optimization has

**Table 6: The classification accuracy of the trained base model on each dataset.**

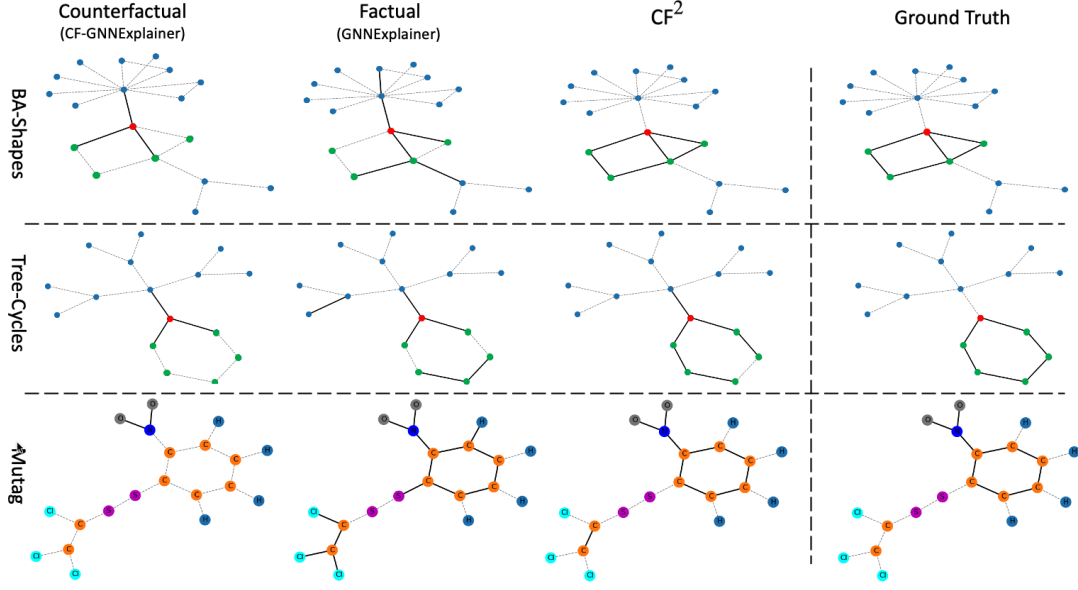
Datasets	BA-Shapes	Tree-Cycles	Mutag <sub>0</sub>	NCI1	CiteSeer
Epochs	3000	3000	1000	200	200
Accuracy	97.86	98.29	98.05	69.03	71.04

more precise prediction but tends to be conservative and low in coverage. Factual-based optimization discovers larger portion of the motifs but also covers redundant edges. In general, CF<sup>2</sup> outperforms the other two methods by considering both necessity and sufficiency in the optimization.

## 7.6 Influence of $\alpha$

The  $\alpha$  in Eq.(12) controls the balance between factual reasoning and counterfactual reasoning. When  $\alpha$  is greater than 0.5, CF<sup>2</sup> considers factual reasoning more than counterfactual reasoning, and when it is less than 0.5, counterfactual reasoning is considered more than factual reasoning. Figure 3 shows the influence of  $\alpha$  on CF<sup>2</sup> when generating explanations for BA-Shapes and Mutag<sub>0</sub> datasets. Result shows that the value of  $\alpha$  is not sensitive, and no matter which





**Figure 2: Qualitative Analysis.** Illustration of the generated explanations on instances from two synthetic datasets, BA-Shapes and Tree-Cycles, and one real-world dataset, Mutag. From left to right, we show the explanations generated by the methods based on counterfactual reasoning (i.e., CF-GNNExplainer), factual reasoning (i.e., GNNExplainer),  $CF^2$ , and ground-truth explanation.

$\alpha$  value we choose in  $(0, 1)$ , the generated explanations are better than only considering one type of reasoning (i.e.,  $\alpha = 0$  or  $\alpha = 1$ ).

### 7.7 Justification of the Evaluation Metric

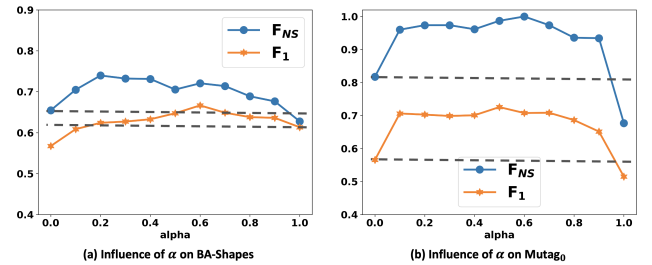
To justify the effectiveness of our PN/PS-based evaluation, we test it on the three datasets with ground-truth explanations, i.e., BA-Shapes, Tree-Cycles, and Mutag<sub>0</sub>. We use two non-parametric methods to test the correlation between the performance on ground-truth evaluation and PN/PS-based evaluation, which are the Kendall's  $\tau$  [18] and Spearman's  $\rho$  [48] scores. These two scores are in the range of  $(-1, 1)$ . Two values are considered positively correlated if  $\tau$  and  $\rho$  are positive scores. The higher the scores are, the closer our proposed evaluation metric is compared to ground-truth evaluation, i.e., can be trusted more to evaluate a given explainable GNN model when the ground-truth is not accessible. We test the correlation between  $F_{NS}$  and  $F_1$ /Accuracy. The results are reported in Table 7. The  $\tau$  and  $\rho$  show that they are highly positively correlated. This is important since for a dataset without ground-truth motifs, if one explanation method performs better than another one according to the PN/PS-based evaluation, then we can have a good confidence to expect the same conclusion if traditional evaluation metrics are used assuming ground-truth is available.

## 8 CONCLUSIONS AND FUTURE WORK

In this work, we propose a Counterfactual and Factual reasoning ( $CF^2$ ) framework, which generates GNN explanations by simultaneously considering the necessity and sufficiency of the explanations. Moreover, we leverage the insights from causal inference theory by taking the Probability of Necessity (PN) and Probability of Sufficiency (PS) to evaluate the necessity and sufficiency of the extracted explanations, making it possible to conduct quantitative evaluation

**Table 7: Correlation between PN/PS-based evaluation and ground-truth evaluation.**

Models	BA-Shapes		Tree-Cycles		Mutag <sub>0</sub>	
	$\tau \uparrow$	$\rho \uparrow$	$\tau \uparrow$	$\rho \uparrow$	$\tau \uparrow$	$\rho \uparrow$
$F_{NS}$ & $F_1$	1.00	1.00	1.00	1.00	1.00	1.00
$F_{NS}$ & Acc	0.66	0.79	1.00	1.00	0.66	0.79



**Figure 3: Influence of  $\alpha$  on (a) BA-Shapes and (b) Mutag<sub>0</sub>.**

of GNN explanations. Experiments on both synthetic and real-world datasets verify the superiority of the proposed method as well as the usefulness of the evaluation metrics. In the future, we will generalize our framework beyond graph-based explanations, including but not limited to vision- and language-based explanations.

## ACKNOWLEDGEMENT

This work was supported in part by NSF IIS 1910154, 2007907, and 2046457. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsors.



## REFERENCES

- [1] Federico Baldassarre and Hossein Azizpour. 2019. Explainability Techniques for Graph Convolutional Networks. In *International Conference on Machine Learning (ICML) Workshops, 2019 Workshop on Learning and Reasoning with Graph-Structured Representations*.
- [2] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K Warmuth. 1987. Occam's razor. *Information processing letters* 24, 6 (1987), 377–380.
- [3] Emmanuel J Candes, Justin K Romberg, and Terence Tao. 2006. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences* 59, 8 (2006), 1207–1223.
- [4] Emmanuel J Candes and Terence Tao. 2005. Decoding by linear programming. *IEEE transactions on information theory* 51, 12 (2005), 4203–4215.
- [5] Hanxiong Chen, Yunqi Li, Shaoyun Shi, Shuchang Liu, He Zhu, and Yongfeng Zhang. 2021. Graph Collaborative Reasoning. *WSDM* (2021).
- [6] Hanxiong Chen, Shaoyun Shi, Yunqi Li, and Yongfeng Zhang. 2021. Neural collaborative reasoning. In *Proceedings of the Web Conference 2021*. 1516–1527.
- [7] Eric Chu, Deb Roy, and Jacob Andreas. 2020. Are visual explanations useful? a case study in model-in-the-loop prediction. *arXiv preprint arXiv:2007.12248* (2020).
- [8] Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kaws, and Prithviraj Sen. 2020. A Survey of the State of Explainable AI for Natural Language Processing. In *AACL*. 447–459.
- [9] Asim Kumar Debnath, 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 (1991), 786–797.
- [10] Hugo Jair Escalante, Sergio Escalera, Isabelle Guyon, Xavier Baró, Yağmur Güçlütürk, Umut Güçlü, Marcel van Gerven, and Rob van Lier. 2018. *Explainable and interpretable models in computer vision and machine learning*. Springer.
- [11] Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. 2018. Pathologies of neural models make interpretations difficult. *EMNLP* (2018).
- [12] Matthew Finlayson, Aaron Mueller, Sebastian Gehrmann, Stuart Shieber, Tal Linzen, and Yonatan Belinkov. 2021. Causal Analysis of Syntactic Agreement Mechanisms in Neural Language Models. In *ACL-IJCNLP*.
- [13] Shijie Geng, Zuohui Fu, Juntao Tan, Yingqiang Ge, Gerard De Melo, and Yongfeng Zhang. 2022. Path Language Modeling over Knowledge Graphs for Explainable Recommendation. *WWW* (2022).
- [14] Lise Getoor. 2005. Link-based classification. In *Advanced methods for knowledge discovery from complex data*. Springer, 189–207.
- [15] Yash Goyal, Ziyang Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Counterfactual visual explanations. In *International Conference on Machine Learning*. PMLR, 2376–2384.
- [16] Braden Hancock, Martin Bringmann, Paroma Varma, Percy Liang, Stephanie Wang, and Christopher Ré. 2018. Training classifiers with natural language explanations. In *Proceedings of the conference. Association for Computational Linguistics Meeting*, Vol. 2018. NIH Public Access, 1884.
- [17] Qiang Huang, Makoto Yamada, Yuan Tian, Dinesh Singh, Dawei Yin, and Yi Chang. 2020. Graphlime: Local interpretable model explanations for graph neural networks. *arXiv preprint arXiv:2001.06216* (2020).
- [18] M. G. Kendall. 1945. The Treatment of ties in ranking problems. *Biometrika* 33, 3 (11 1945), 239–251. <https://doi.org/10.1093/biomet/33.3.239> [arXiv:https://academic.oup.com/biomet/article-pdf/33/3/239/573257/33-3-239.pdf](https://academic.oup.com/biomet/article-pdf/33/3/239/573257/33-3-239.pdf)
- [19] Jiwei Li, Will Monroe, and Dan Jurafsky. 2016. Understanding neural networks through representation erasure. *arXiv preprint arXiv:1612.08220* (2016).
- [20] Lei Li, Yongfeng Zhang, and Li Chen. 2021. Personalized Transformer for Explainable Recommendation. In *ACL*. 4947–4957.
- [21] Chris Lin, Gerald J Sun, Krishna C Bulusu, Jonathan R Dry, and Marylens Hernandez. 2020. Graph neural networks including sparse interpretability. *arXiv preprint arXiv:2007.00119* (2020).
- [22] Wanyu Lin, Hao Lan, and Baochun Li. 2021. Generative Causal Explanations for Graph Neural Networks. In *Proceedings of the 38th International Conference on Machine Learning*. 6666–6679.
- [23] Ana Lucic, Maartje ter Hoeve, Gabriele Tolomei, Maarten de Rijke, and Fabrizio Silvestri. 2021. CF-GNNExpainer: counterfactual explanations for graph neural networks. *arXiv:2102.03322* (2021).
- [24] Dongsheng Luo, Wei Cheng, Dongkuan Xu, Wenchao Yu, Bo Zong, Haifeng Chen, and Xiang Zhang. 2020. Parameterized Explainer for Graph Neural Network. In *Advances in Neural Information Processing Systems*.
- [25] Oisín Mac Aodha, Shihan Su, Yuxin Chen, Pietro Perona, and Yisong Yue. 2018. Teaching categories to human learners with visual explanations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3820–3828.
- [26] Andreas Madsen, Siva Reddy, and Sarath Chandar. 2021. Post-hoc Interpretability for Neural NLP: A Survey. *arXiv preprint arXiv:2108.04840* (2021).
- [27] Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. 2021. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12700–12710.
- [28] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. 2016. *Causal inference in statistics: A primer*. John Wiley & Sons.
- [29] Phillip E Pope, Soheil Kolouri, Mohammad Rostami, Charles E Martin, and Heiko Hoffmann. 2019. Explainability methods for graph convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10772–10781.
- [30] Ana Cristina Quelhas, Célia Rasga, and P. N. Johnson-Laird. 2018. The Relation Between Factual and Counterfactual Conditionals. *Cognitive Science* 42, 7 (2018), 2205–2228.
- [31] 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.
- [32] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*. 618–626.
- [33] Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Galligher, and Tina Eliassi-Rad. 2008. Collective classification in network data. *AI magazine* 29, 3 (2008), 93–93.
- [34] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In *International Conference on Machine Learning*. PMLR, 3145–3153.
- [35] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. In *Workshop at International Conference on Learning Representations*.
- [36] Juntao Tan, Shuyuan Xu, Yingqiang Ge, Yunqi Li, Xu Chen, and Yongfeng Zhang. 2021. Counterfactual Explainable Recommendation. *CIKM* (2021).
- [37] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. 2020. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3716–3725.
- [38] Tom Vermeire, Dieter Brughmans, Sofie Goethals, Raphael Mazzine Barbosa de Oliveira, and David Martens. 2022. Explainable image classification with evidence counterfactual. *Pattern Analysis and Applications* (2022), 1–21.
- [39] Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart M Shieber. 2020. Investigating Gender Bias in Language Models Using Causal Mediation Analysis. In *Advances in Neural Information Processing Systems*.
- [40] Nikil Wale and George Karypis. 2006. Comparison of Descriptor Spaces for Chemical Compound Retrieval and Classification. In *International Conference on Data Mining*. 678–689.
- [41] Yikun Xian, Zuohui Fu, Shan Muthukrishnan, Gerard De Melo, and Yongfeng Zhang. 2019. Reinforcement knowledge graph reasoning for explainable recommendation. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*. 285–294.
- [42] Pinar Yanardag and SVN Vishwanathan. 2015. Deep graph kernels. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. 1365–1374.
- [43] Zhitao Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. 2019. GNNExplainer: Generating Explanations for Graph Neural Networks. In *Advances in Neural Information Processing Systems*.
- [44] Hao Yuan, Jiliang Tang, Xia Hu, and Shuiwang Ji. 2020. Xggn: Towards model-level explanations of graph neural networks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 430–438.
- [45] Hao Yuan, Haiyang Yu, Shurui Gui, and Shuiwang Ji. 2020. Explainability in graph neural networks: A taxonomic survey. *arXiv preprint arXiv:2012.15445* (2020).
- [46] Yongfeng Zhang and Xu Chen. 2020. Explainable recommendation: A survey and new perspectives. *Foundations and Trends in Information Retrieval* (2020).
- [47] Yongfeng Zhang, Guokun Lai, Min Zhang, Yi Zhang, Yiqun Liu, and Shaoping Ma. 2014. Explicit factor models for explainable recommendation based on phrase-level sentiment analysis. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. 83–92.
- [48] Daniel Zwillinger and Stephen Kokoska. 1999. *CRC standard probability and statistics tables and formulae*. Crc Press.

## A MATHEMATICAL DEFINITIONS FOR NODE CLASSIFICATION

In Section 4 and Section 5, we formulate the Explainable GNN problem as well as CF<sup>2</sup> framework, under the graph classification setting. In this section, we provide the same mathematical definition under node classification task.

### A.1 Problem Formulation (Node Classification)

**Explainable Graph Neural Networks** In a given graph  $G = \{\mathcal{V}, \mathcal{E}\}$ , Suppose a node  $v_i \in G$  has the predicted label  $\hat{y}_i$ . The computational graph for node  $v_i$  is defined as  $G_{s(i)} = \{\mathcal{V}_{s(i)}, \mathcal{E}_{s(i)}\}$ , which is a sub-graph of  $G$  that consists of the  $L$ -hop neighbors of node  $v_i$ .  $A_{s(i)} \in \{0, 1\}^{|\mathcal{V}_{s(i)}| \times |\mathcal{V}_{s(i)}|}$  and  $X_{s(i)} \in \mathbb{R}^{|\mathcal{V}_{s(i)}| \times d}$  are the related adjacency matrix and feature matrix of the computational graph. Since only  $G_{s(i)}$  will influence the prediction made by the GNN model, the generated explanation should be a sub-graph of  $G_{s(i)}$ . Thus, for node classification task, the goal of the explainable GNN problem is to learn an edge mask  $M_{s(i)} \in \{0, 1\}^{|\mathcal{V}_{s(i)}| \times |\mathcal{V}_{s(i)}|}$  and a feature mask  $F_{s(i)} \in \{0, 1\}^{|\mathcal{V}_{s(i)}| \times d}$ , which will be applied on  $A_{s(i)}$  and  $X_{s(i)}$ , respectively. After optimization, the sub-graph will be  $A_{s(i)} \odot M_{s(i)}$  with the sub-features  $X_{s(i)} \odot F_{s(i)}$ , which is the generated explanation for the prediction of node  $v_i$ .

**Counterfactual and Factual Conditions** For node classification task, the definition of the conditions for factual and counterfactual is similar to graph classification, which are defined as following:

Condition for Factual Reasoning :

$$\arg \max_{c \in C} P_{\Phi}(c \mid A_{s(i)} \odot M_{s(i)}, X_{s(i)} \odot F_{s(i)}) = \hat{y}_i \quad (15)$$

Condition for Counterfactual Reasoning :

$$\arg \max_{c \in C} P_{\Phi}(c \mid A_{s(i)} - A_{s(i)} \odot M_{s(i)}, X_{s(i)} - X_{s(i)} \odot F_{s(i)}) \neq \hat{y}_i \quad (16)$$

**Simple and Effective Explanations** For node classification task, the explanation complexity is defined exactly the same to graph classification, which is:

$$C(M, F) = \|M\|_0 + \|F\|_0 \quad (17)$$

The factual explanation strength and counterfactual explanation strength are defined with the node classification settings as:

$$S_f(M, F) = P_{\Phi}(\hat{y}_{i,s} \mid A_{s(i)} \odot M_{s(i)}, X_{s(i)} \odot F_{s(i)}) \quad (18)$$

and

$$S_c(M, F) = -P_{\Phi}(\hat{y}_{i,s} \mid A_{s(i)} - A_{s(i)} \odot M_{s(i)}, X_{s(i)} - X_{s(i)} \odot F_{s(i)}) \quad (19)$$

### A.2 The CF<sup>2</sup> Framework (Node Classification)

The basic idea of CF<sup>2</sup> for node and graph classification are same, which is minimizing the explanation complexity while the generated explanation is strong enough. Therefore, we directly provide the final relaxed optimization and omit the derivation process. CF<sup>2</sup> generates explanations via solving the relaxed optimization equation:

$$\text{minimize } \|M_{s(i)}^*\|_1 + \|F_{s(i)}^*\|_1 + \lambda(\alpha L_f + (1 - \alpha)L_c) \quad (20)$$

where

$$L_f = \text{ReLU}(\gamma + P_{\Phi}(\hat{y}_{i,s} \mid A_{s(i)} \odot M_{s(i)}^*, X_{s(i)} \odot F_{s(i)}^*) - S_f(M_{s(i)}^*, F_{s(i)}^*)) \quad (21)$$

Similarly,

$$L_c = \text{ReLU}(\gamma - S_c(M_{s(i)}^*, F_{s(i)}^*) - P_{\Phi}(\hat{y}_{i,s} \mid A_{s(i)} - A_{s(i)} \odot M_{s(i)}^*, X_{s(i)} - X_{s(i)} \odot F_{s(i)}^*)) \quad (22)$$