Zorro

Valid, sparse, and stable explanations in graph neural networks

Funke, Thorben; Khosla, Megha; Rathee, Mandeep; Anand, Avishek

DOI
10.1109/TKDE.2022.3201170

Publication date
2023

Document Version
Final published version

Published in
IEEE Transactions on Knowledge & Data Engineering

Citation (APA)
Funke, T., Khosla, M., Rathee, M., & Anand, A. (2023). Zorro: Valid, sparse, and stable explanations in graph neural networks. *IEEE Transactions on Knowledge & Data Engineering*, *35*(8), 8687-8698. Article 9866587. https://doi.org/10.1109/TKDE.2022.3201170

ZORRO: Valid, Sparse, and Stable Explanations in Graph Neural Networks

Thorben Funke, Megha Khosla, Mandeep Rathee, and Avishek Anand

Abstract—With the ever-increasing popularity and applications of graph neural networks, several proposals have been made to explain and understand the decisions of a graph neural network. Explanations for graph neural networks differ in principle from other input settings. It is important to attribute the decision to input features and other related instances connected by the graph structure. We find that the previous explanation generation approaches that maximize the mutual information between the label distribution produced by the model and the explanation to be restrictive. Specifically, existing approaches do not enforce explanations to be valid, sparse, or robust to input perturbations. In this paper, we lay down some of the fundamental principles that an explanation method for graph neural networks should follow and introduce a metric *RDT-Fidelity* as a measure of the explanation's effectiveness. We propose a novel approach Zorro based on the principles from *rate-distortion theory* that uses a simple combinatorial procedure to optimize for RDT-Fidelity. Extensive experiments on real and synthetic datasets reveal that Zorro produces sparser, stable, and more faithful explanations than existing graph neural network explanation approaches.

Index Terms—Explainability, graph neural networks, interpretability

1 INTRODUCTION

GRAPH neural networks (GNNs) are a flexible and powerful family of models that build representations of nodes or edges on irregular graph-structured data and have experienced significant attention in recent years. GNNs are based on the "neighborhood aggregation" scheme, where a node representation is learned by aggregating features from their neighbors. Learning complex neighborhood aggregations and latent feature extraction has enabled GNNs to achieve state-of-the-art performance on node and graph classification tasks. This complexity, on the other hand, leads to a more opaque and non-interpretable model. To alleviate the problem of interpretability, we focus on explaining the rationale underlying a given prediction of already trained GNNs.

There are diverse notions and regimes of explainability and interpretability for machine learning models – (a) interpretable models versus post-hoc explanations, (b) model-introspective versus model-agnostic explanations, (c) outputs in terms of feature versus data attributions [23], [28]. In this work, we aim to explain the decision of an already trained GNN model, i.e., *compute post-hoc explanations for a trained GNN*. Additionally, we do not assume any access to the trained model parameters, i.e., we are model-agnostic or black-box regime. Finally, our explanation attributes the reason for an underlying GNN prediction to either a subset of features or neighboring nodes or both.

There has been recent interest in designing explainers for GNNs that produce feature attributions in a post-hoc manner [30], [40], [44] where a combination of nodes, edges, or features is retrieved as an explanation. We introduce some essential notions of validity, sparsity, and stability for explaining GNNs and argue that many of the existing works on explainable GNNs do not satisfy these principles. To systematically fill in the above gaps, we commence by formulating three desired properties of a GNN explanation: *validity*, *sparsity*, and *stability*. Fig. 1 provides an illustration of these three properties.

Validity. Existing explanations approaches used for explaining GNNs like gradient-based feature attribution techniques select nodes or features [38] are not optimized to be valid as well as being explanatory. An explanation is valid if just the explanation (a subset of features and nodes) as input would be sufficient to arrive at the same prediction.

Sparsity. It is easy to see that validity alone is not sufficient for an explanation as the entire input is a valid explanation [41]. Ideally, the explanation should only highlight those parts of the input with the highest discriminative information. Existing explanation approaches accomplish this by outputting distributions or *soft-masks* over input features or nodes [44]. However, humans find it hard to make sense of soft masks and instead prefer sparse binary masks or *hard masks* [1], [19], [27], [49]. We define *sparsity* as the *size of the explanation in terms of number of non-zero elements* in the explanation. A sparse explanation in the form of a hard mask is, therefore, more desirable and reduces ambiguities due to soft masks [15].

- *Thorben Funke and Mandeep Rathee are with the L3S Research Center, Leibniz University Hanover, 30167 Hannover, Germany. E-mail: {tfunke, rathee}@l3s.de.*
- *Megha Khosla and Avishek Anand are with the TU Delft, 2628, CD, Delft, Netherlands. E-mail: M.Khosla@tudelft.nl, anand@l3s.de.*

Manuscript received 29 July 2021; revised 30 June 2022; accepted 6 August 2022. Date of publication 24 August 2022; date of current version 21 June 2023.
This work was partially supported in part by the project "CampaNeo" under Grant ID 01MD19007A funded in part by the BMWi, and the European Commission EU H2020, "smashHit", under Grant 871477.
(Corresponding author: Thorben Funke.)
Recommended for acceptance by Y. Xia.
This article has supplementary downloadable material available at https://doi.org/10.1109/TKDE.2022.3201170, provided by the authors.
Digital Object Identifier no. 10.1109/TKDE.2022.3201170

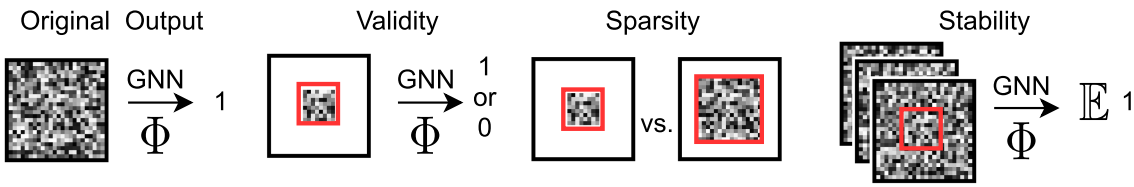

Fig. 1. Illustration of validity, sparsity, and stability. The GNN Φ takes the feature matrix X, which is illustrated as a grayscale matrix, and the relations from the graph G, which is not shown for simplicity, to predict the class label (1). An explanation selects the most important inputs from the feature matrix responsible for the prediction, which we illustrate as red rectangles. The validity of an explanation is the property to preserve the prediction if a fixed baseline value replaces all not selected values. The sparsity is the number of selected elements, where fewer elements are desirable. Lastly, stability is the property to preserve the prediction if all not selected values are perturbed. Existing methods only optimize for validity and sparsity. However, even trivial explanations can be valid and sparse.

Stability. Validity and sparsity, though necessary, are not sufficient to define an explanation. In Section 3.1, we show that the trivial empty explanation (all features are replaced by 0s) could be a valid explanation for many cases. The high validity observed in such cases is an artifact of a particular configuration of trained model parameters. We would rather expect that the model retains its predicted class with only the knowledge of the explanation *while the rest of the information in the input is filled randomly*. In other words, an explanation should be valid independent of the rest of the input. We say that an explanation is *stable* if the behavior of the GNN is unaffected by the features outside of the explanation. Most of the existing works do not consider stability in their modeling of explanation approaches.

In this article, we introduce a new metric called *RDT-Fidelity* which is grounded in the principles of *rate-distortion theory* and reflects these three desiderata into a single measure. Essentially, we cast the problem of finding explanations given a trained model as a signal/message reconstruction task involving a sender, a receiver, and a noisy channel. The message sent by the sender is the actual feature vector, with the explanation being a subset of immutable feature values. The noisy channel can obfuscate only the features that do not belong to the explanation. The explanation's RDT-Fidelity lies in the degree to which the decoder can faithfully reconstruct from the noisy feature vector. Maximizing RDT-Fidelity while ensuring explanation sparsity is NP-hard (for justification refer to Section 5), and we consequently propose a greedy combinatorial procedure Zorro that generates sparse, valid, and stable explanations.

Accurately measuring the effectiveness of post-hoc explanations has been acknowledged to be a challenging problem due to the lack of explanation ground truth. We carry out an extensive and comprehensive experimental study on several experimental regimes [12], [30], [44], three real-world datasets [43] and four different GNN architectures [20], [21], [39], [42] to evaluate the effectiveness of our explanations. In addition to measuring validity, sparsity, and RDT-Fidelity, we also compare our approach with the evaluation regime proposed in GNNExplainer [44], the faithfulness measure proposed in [30] and the ROAR methodology from [12].

First, we establish that Zorro outperforms all other baselines over different evaluation regimes on both real-world and synthetic datasets. Based on Zorro's explanation, we retrieve valuable insights into the GNN's behavior: different GNN's derive their decisions from different large portions of the input, and more available features do not mean more relevant features; the GNN's base their classification on different scales on the local homophily; multiple disjoint explanations are possible, i.e., GNN's classification is derived from disjoint parts of the network (duplicated information flow).

To sum up, our main contributions are:

- We theoretically investigate the key properties of *validity*, *sparsity*, and *stability* that a GNN explanation should follow.
- We introduce a novel evaluation metric, *RDT-Fidelity* derived from principles of *rate-distortion theory* that reflects these desiderata into a single measure.
- We propose a simple combinatorial called Zorro to find high RDT-Fidelity explanations with theoretically bounded stability. We *release our code* at https://github.com/funket/zorro.
- We perform extensive scale experiments on synthetic and real-world datasets. We show that Zorro not only outperforms baselines for RDT-Fidelity but also for several evaluation regimes so far proposed in the literature.

2 Related Work

Representation learning approaches on graphs encode graph structure into low-dimensional vector representations, using deep learning and nonlinear dimensionality reduction techniques. These representations are trained in an unsupervised [10], [18], [25] or semi-supervised manner by using neighborhood aggregation strategies and task-based objectives [20], [39].

2.1 Explainability in Machine Learning

Post-hoc approaches to model explainabiliy are popularized by *feature attribution* methods that aim to assign importance to input features given a prediction either agnostic to the model parameters [28], [29] or using model specific attribution approaches [3], [38]. *Instance-wise feature selection* (IFS) approaches [7], [45], on the other hand, focuses on finding a *sufficient* feature subset or explanation that leads to little or no degradation of the prediction accuracy when other features are masked. Applying these works directly for graph models is infeasible due to the complex form of explanation, which should consider the complex association among nodes and input features.

2.2 Explainability in GNNs

Explainability approaches for explaining node level decisions include soft-masking approaches [11], [22], [24], [31], [32], [44], Shapely based approaches [8], [48], surrogate model based methods [13], [40], and gradients based methods [17], [26], [30]. Soft-masking approaches like GNNExplainer [44] learns a real-valued edge and feature mask such that the mutual information with GNN's predictions is maximized.

An example of a surrogate model based method is PGMExplainer [40] which builds a simpler interpretable Bayesian network explaining the GNN prediction. Others adopt existing explanations approaches such as Shapely [8], [48], layer-wise relevance propagation [32], causal effects [22] or LIME [13], [17], to graph data.

The key idea in gradient based methods is to use the gradients or hidden feature map values to approximate input importance. This approach is the most straightforward solution to explain deep models and is quite popular for image and text data. For graph data, [26] and [30] applied gradient based methods for explaining GNNs, which rely on propagating gradients/relevance from the output to the original model's input.

Another line of work that focuses on explaining decisions at a graph level includes XGNN [46] and GNES [11]. XGNN proposed a reinforcement learning-based graph generation approach to generate explanations for the predicted class for a graph. GNES jointly optimizes task prediction and model explanation by enforcing graph regularization and weak supervision on model explanations. Other works [14], [16] focus on explaining unsupervised network representations, which is out of scope for the current work or are specific to the combination of GNNs and NLP [31].

Most of the existing approaches for explaining GNNs are based on soft-masking methods [11], [22], [24], [26], [32], [44]. However, soft masks are typically hard for humans to interpret than hard masks due to their low sparsity and inherent uncertainty [1], [19], [27], [49]. Only a few hard-masking approaches for GNNs exist. PGMExplainer [40] defines explanation in terms of relevant neighborhood nodes influencing the model decision and does not consider node features. PGExplainer [24] employs a parameterized model to generate soft edge masks with node representations (extracted from target GNN) as input. Unlike our approach, PGExplainer is not model agnostic. Like PGMExplainer, it also does not generate a feature-based explanation. SubgraphX [48] optimizes for Shapely values based on a Monte Carlo tree search.

3 Properties of GNN Explanations

3.1 Defining GNN Explanation, Validity and Sparsity

We are interested in explaining the prediction of the GNN $\Phi(n)$ for any node n. Specifically, we consider the task of node classification. We note that for a particular node n, the subgraph taking part in the computation of neighborhood aggregation operation, see Eq. (5), fully determines the information used by GNN to predict its class. In particular, for a L-layer GNN, this subgraph would be the graph induced on nodes in the L-hop neighborhood of n. For brevity, we call this subgraph the *computational graph* of the query node. We want to point out that the term "computational graph" should not be confused with the neural network's computational graph. For a brief description of GNNs in general, see Appendix A, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TKDE.2022.3201170.

Let $G(n) \subseteq G$ denote the computational graph of the node n. Let $X(n)$, or briefly X denotes the feature matrix restricted to the nodes of $G(n)$, where each row corresponds to a d-dimensional feature vector of the corresponding node in the computational graph. We define explanation $\mathcal{S} = \{F_s, V_s\}$ as a subset of input features and nodes. In principle, $\mathcal{S}$ would correspond to the feature matrix restricted to features in F_s of nodes in V_s. We quantify the validity and sparsity of $\mathcal{S}$ as follows.

Definition 1 (Validity). *The validity score of explanation $\mathcal{S}$ is 1 if $\Phi(\mathcal{S}) = \Phi(X)$ and 0 otherwise.*

In literature, the validity of an explanation is usually computed with respect to the baseline 0, i.e., we set values of all features not in $\mathcal{S}$ to 0. An alternative is to use the average of the feature scores instead. As we discuss in Appendix D.3, available in the online supplemental material, our validity is related to one of the metrics from [36], [37], [47].

Definition 2 (Sparsity). *The sparsity of an explanation is measured as the ratio of bits required to encode an explanation to those required to encode the input. We use explanation entropy to compare sparsity for a fixed input and call this the effective explanation size.*

In contrast to other sparsity definitions, such as in [47] our definition of sparsity is more general. It can be directly applied for both hard-masks and soft-masks without the need for any transformation. Without loss of generality, we can assume that an explanation is a continuous mask over the set of features and nodes/edges where the mask value quantifies the importance of the corresponding element. We state the upper bound of the sparsity value in the following proposition.

Proposition 1. *Let p be the normalized distribution of explanation (feature) masks. Then sparsity of an explanation is given by the entropy $H(p)$ and is bounded from above by $\log(|M|)$ where M corresponds to a complete set of features or nodes.*

Proof. We first compute the normalized feature or node mask distribution, $p(f)$ for $f \in M$. In particular, denoting the mask value of f by $\mathrm{mask}(f)$, we have

$$p(f) = \frac{\mathrm{mask}(f)}{\sum_{f' \in M} \mathrm{mask}(f')}.$$

Then $H(p) = -\sum_{f \in M} p(f) \log p(f)$ which achieves its maximum for the uniform distribution, i.e., $p(f) = \frac{1}{|M|}$. □

3.2 Limitations of Validity and Sparsity

We illustrate the limitations of previous works, which are based on maximizing validity and sparsity of explanations by a simple example shown in Fig. 2. The example is inspired by the example for text analysis from [4].

The input is a graph with node set $V = \{v_1, v_2, v_3, v_4\}$. Each node has a single feature, with the value given in the Figure.

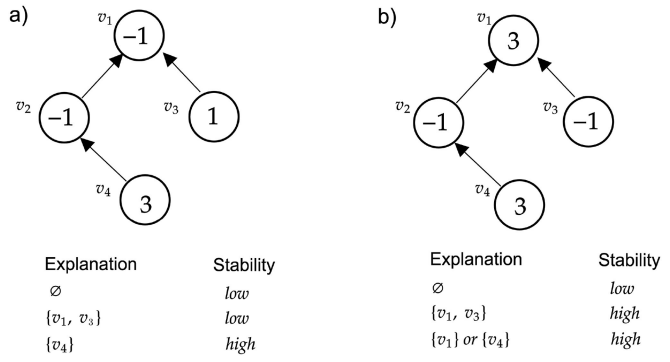

Fig. 2. In this synthetic example, we approximate the (node) classification of v_1 by GNN with a rule based on the sum of the node features $\sum f(v_i)$. All given explanations are valid (when the unselected input is set to 0) and sparse. However, we see that in a) the explanation $\{v_1, v_3\}$ has the same stability as the trivial mask. Example b) highlights that selecting additional elements may not decrease the stability and that even two disjoint explanations are possible.

Let us assume that these feature values lie in the range from -2 to 3. For any node v, we define the model output in terms of simple sum (aggregation) of feature values, $f(v)$ of itself and its two hop-neighborhood. For example,

$$\Phi(v_1) = \begin{cases} 1, & \text{if } f(v_1) + f(v_2) + f(v_3) + f(v_4) \geq 0, \\ 0, & \text{otherwise.} \end{cases}$$

Now we wish to explain the prediction $\Phi(v_1) = 1$.

Consider an explanation $\{v_1, v_3\}$. Clearly it is valid explanation with a validity score of 1, if we set the not selected nodes' features to 0. But if we set $f(v_4) = -2, f(v_2) = 0$, the explanation $\{v_1, v_3\}$ is no longer valid.

Similarly, the empty explanation, $\mathcal{S} = \emptyset$, is the sparsest possible explanation which has a validity score of 1 when all feature values are set to 0. However, for a different realization of the unimportant features, say for $f(v_2) = -1$ and rest all are set to 0, the validity score is reduced to 0.

We want to emphasize that a particular explanation $\{v_1, v_3\}$ or an empty explanation might be valid for an individual configuration of the features of not selected vertices but not for others. However, a proper explanation should explain the model's prediction independent of the remaining input configuration.

This subtle point is usually ignored by existing explainability approaches, which only evaluate an explanation for a specific baseline of the irrelevant part of the input. In contrast, we propose stability, which takes into account the variance of the validity of an explanation over different configurations of the input's unselected parts.

Definition 3 (Stability). *Let $\mathcal{Y}$ be a random variable sampled from the distribution over validity scores for different realizations of $X \setminus \mathcal{S}$. Let $\mathrm{Var}(\mathcal{Y})$ denote the variance of $\mathcal{Y}$. We define stability $\gamma(\mathcal{S})$ of an explanation, $\mathcal{S}$ as*

$$\gamma(\mathcal{S}) = \frac{1}{1 + \mathrm{Var}(\mathcal{Y})}.$$

Note that $\gamma(\mathcal{S}) \in (0, 1]$ holds and achieves, on the one hand, maximum value of 1 if $\mathrm{Var}(\mathcal{Y}) = 0$, i.e., when the explanation is completely independent of components in $X \setminus \mathcal{S}$. On the other hand, the stability will also be equal to 1 if the validity of an explanation for all realizations is equal to zero. Mathematically, we need to ensure a high expected value of $\mathcal{Y}$ in addition to its low variance. Therefore, we need another metric along with stability.

To account for the stability of explanations, we introduce a novel metric called *RDT-Fidelity* which has a sound theoretical grounding in the area of *rate-distortion theory* [35]. We describe RDT-Fidelity and its relation to rate-distortion theory and stability in the next section.

4 Rate-Distortion Theory and RDT-Fidelity

Rate-distortion theory addresses the problem of determining the minimum number of bits per symbol (also referred to as *rate*) that should be communicated over a channel so that the source signal can be approximately reconstructed at the receiver without exceeding an expected distortion, D. Mathematically we are interested in finding the conditional probability density function, $Q(\mathcal{S}|X)$, of the compressed signal or explanation S given the input X such that the expected distortion $D(S, X)$ is upper bounded.

$$\inf_{Q(\mathcal{S}|X)} I_Q(\mathcal{S}, X) \text{ such that } \mathbb{E}_Q(D(\mathcal{S}, X)) \leq D^*, \tag{1}$$

where $I(\mathcal{S}, X)$ denotes the mutual information between input X and compressed signal S and D^* corresponds to maximum allowed distortion. Note that Eq. (1) requires minimization of mutual information between X and $\mathcal{S}$. Mutual information will be minimized when $\mathcal{S}$ is completely independent of X.

In our explanation framework, the compressed signal $\mathcal{S}$ corresponds to an explanation. The effect of minimizing the mutual information between the compressed signal and the input, see Eq. (1), would amount to minimize the size of $\mathcal{S}$. A trivial solution of the empty set is avoided by restricting the average distortion of $\mathcal{S}$ in the second part of the objective. In particular, compressed signal (explanation) should be such that knowing only about the input on $\mathcal{S}$ and filling in the rest of the information randomly will almost surely preserve the desired output signal (or class prediction).

In particular, for graph models, the explanation $\mathcal{S}$ which is a subset of input nodes as well as input features, is most relevant for a classification decision if the expected classifier score remains nearly the same when randomizing the remaining input $X \setminus \mathcal{S}$.

More precisely, we formulate the task of explaining the model prediction for a node n, as finding a partition of the components of its computational graph into a subset, $\mathcal{S}$ of relevant nodes and features, and its complement $\mathcal{S}^c$ of non-relevant components. In particular, the subset $\mathcal{S}$ should be such that fixing its value to the true values already determines the model output for almost all possible assignments to the non-relevant subset $\mathcal{S}^c$. The subset $\mathcal{S}$ is then returned as an explanation. As it is a rate-distortion framework, we are interested in an explanation (compressed signal) with the maximum agreement (minimum distortion) with the actual model's prediction on complete input. This agreement, what we refer to as *RDT-Fidelity* is quantified by the expected validity score of an explanation over all possible configurations of the complement set $\mathcal{S}^c$.

4.1 RDT-Fidelity

To formally define RDT-Fidelity, let us denote with $Y_{\mathcal{S}}$ the perturbed feature matrix obtained by fixing the components of the $\mathcal{S}$ to their actual values and otherwise noisy entries. The values of components in $\mathcal{S}^c$ are then drawn from some noisy distribution, $\mathcal{N}$. Let $\mathcal{S} = \{V_s, F_s\}$ be the explanation with selected nodes V_s and selected features F_s.

Let $M(\mathcal{S})$, or briefly M, be the mask matrix such that each element $M_{i,j} = 1$ if and only if ith node (in $G(n)$) and jth feature are included in sets V_s and F_s respectively and 0 otherwise. Then the perturbed input is given by

$$Y_{\mathcal{S}} = X \odot M(\mathcal{S}) + Z \odot (\mathbb{1} - M(\mathcal{S})), Z \sim \mathcal{N}, \tag{2}$$

where $\odot$ denotes an element-wise multiplication, and $\mathbb{1}$ a matrix of ones with the corresponding size.

Definition 4 (RDT-Fidelity). *The* RDT-Fidelity *of explanation $\mathcal{S}$ with respect to the GNN Φ and the noise distribution $\mathcal{N}$ is given by*

$$\mathcal{F}(\mathcal{S}) = \mathbb{E}_{Y_{\mathcal{S}}|Z\sim\mathcal{N}}\big[\mathbb{1}_{\Phi(X)=\Phi(Y_{\mathcal{S}})}\big]. \tag{3}$$

In simple words, RDT-Fidelity is computed as the expected validity of the perturbed input $Y_{\mathcal{S}}$.

Note that high RDT-Fidelity explanations would be stable by definition, i.e., their validity score would not vary significantly across different realizations of $\mathcal{S}^c$.

Theorem 1. *An explanation with RDT-Fidelity p has stability value of $\frac{1}{1+p(1-p)}$.*

Proof. Let $\mathcal{Y}$ be the random variable corresponding to validity score for an explanation $\mathcal{S}$. Note that

$$\mathbb{E}(\mathcal{Y}) = \mathbb{E}_{Y_{\mathcal{S}}|Z\sim\mathcal{N}}\big[\mathbb{1}_{\Phi(X)=\Phi(Y_{\mathcal{S}})}\big] = p,$$

where Y_S and Z are as defined in Equation (2).

Note that $\mathcal{Y}$ can be understood as a sample drawn from a Bernoulli distribution with a mean equal to RDT-Fidelity value, i.e., $\mathcal{Y} \sim \mathrm{Ber}(p)$ The variance of a Bernoulli distributed variable $\mathcal{Y}$ is given by $p(1-p)$. The proof is completed then by substituting the variance in the definition of stability. □

Theorem 1 implies that for RDT-Fidelity greater than 0.5, the stability increases with an increase in RDT-Fidelity and achieves a maximum value of 1 when RDT-Fidelity reaches its maximum value of 1. Therefore, to ensure high stability, it suffices to find high RDT-Fidelity explanations. As stability is theoretically bounded with respect to RDT-Fidelity, we do not additionally report stability in our experiments. Note that to find non-trivial explanations, we need to maximize RDT-Fidelity together with the sparsity constraint on the explanations. This constraint would, in turn, make the optimization problem NP-Hard. Consequently, we propose a greedy solution as described in the next section.

5 MAXIMIZING RDT-FIDELITY

We propose a simple but effective greedy combinatorial approach, which we call ZORRO, to find high RDT-Fidelity explanations. By fixing the RDT-Fidelity to a certain user-defined threshold, say τ, we are interested in the sparsest explanation, which has a RDT-Fidelity of at least τ. In particular, the problem of finding the sparsest explanation now reduces to finding a minimum subset of features and nodes with RDT-Fidelity of at least τ. It has already been shown in [6] that the problem of selecting the minimum feature subset is NP-Hard.

The pseudocode is provided in Algorithm 1. Let for any node n, V_n denote the vertices in its computational graph $G(n)$, i.e., the set of vertices in L-hop neighborhood of node n for an L-layer GNN; and F denote the complete set of features. We start with zero-sized explanations and select as first element

$$\underset{f\in F}{\operatorname{argmax}}\ \mathcal{F}(V_n, \{f\}) \quad \text{or} \quad \underset{v\in V_n}{\operatorname{argmax}}\ \mathcal{F}(\{v\}, F), \tag{4}$$

whichever yields the highest RDT-Fidelity value. We iteratively add new features or nodes to the explanation such that the RDT-Fidelity is maximized over all evaluated choices. Let V_p and F_p respectively denote the set of possible candidate nodes and features that can be included in an explanation at any iteration. We save for each possible node $v \in V_p$ and feature $f \in F_p$ the ordering R_{V_p} and R_{F_p} given by the RDT-Fidelity values $\mathcal{F}(\{v\}, F_p)$ and $\mathcal{F}(V_p, \{f\})$ respectively. To reduce the computational cost, we only evaluate each iteration the top K remaining nodes and features determined by R_{V_p} and R_{F_p}.

Algorithm 1. ZORRO (n, τ)

Input: node n, threshold τ
Output: explanation, i.e., node mask V_s & feature mask F_s

$V_n \leftarrow$ set of vertices in $G(n)$
$F_p \leftarrow$ set of node features
$V_r = V_p$, $F_r = F_p$, $V_s = \emptyset$, $F_s = \emptyset$
$R_{V_p} \leftarrow$ list of $v \in V_p$ sorted by $\mathcal{F}(\{v\}, F_p)$
$R_{F_p} \leftarrow$ list of $f \in F_p$ sorted by $\mathcal{F}(V_p, \{f\})$
Add maximal element to V_s or F_s as in (4)
while $\mathcal{F}(V_s, F_s) \leq \tau$ **do**
  $\tilde{V}_s = V_s \cup \underset{v\in\mathrm{top}_K(V_r)}{\operatorname{argmax}}\ \mathcal{F}(\{v\} \cup V_s, F_s)$
  $\tilde{F}_s = F_s \cup \underset{f\in\mathrm{top}_K(F_r)}{\operatorname{argmax}}\ \mathcal{F}(V_s, \{f\} \cup F_s)$
  if $\mathcal{F}(\tilde{V}_s, F_s) \leq \mathcal{F}(V_s, \tilde{F}_s)$ **then**
    $F_r = F_r \setminus \{f\}$, $F_s = \tilde{F}_s$
  else
    $V_r = V_r \setminus \{v\}$, $V_s = \tilde{V}_s$
return $\{V_s, F_s\}$

As shown in Fig. 2, an instance can have multiple valid, sparse, and stable explanations. Therefore, we also propose a variant of ZORRO, which continues searching for further explanations: Once we found an explanation with the desired RDT-Fidelity, we discard the chosen elements from the feature matrix X, i.e., we never consider them again as possible choices in computing the next explanation. We repeat the process by finding relevant selections disjoint from the ones already found. To ensure that disjoint elements of the feature matrix X are selected, we recursively call Algorithm 1 with either remaining (not yet selected in any explanation) set of nodes or features. Finally, we return the set of explanations such that the RDT-Fidelity of τ cannot be reached by using all the remaining components that are not in any explanation.

For a detailed explanation of the details and the reasoning behind various design choices, we refer to Appendix C, available in the online supplemental material.

The pseudocode to compute RDT-Fidelity is provided in Algorithm 2. Specifically we generate the obfuscated instance for a given explanation $\mathcal{S} = \{V_s, F_s\}$, $Y_{\mathcal{S}}$ by setting the feature values for selected node-set V_s corresponding to selected features in F_s to their true values. To set the irrelevant values, we randomly choose a value from the set of all possible values for that particular feature in the dataset $\mathcal{X}$. To approximate the expected value in Eq. (3), we generate a finite number of samples of $Y_{\mathcal{S}}$. We then compute RDT-Fidelity as average validity with respect to these different baselines.

Algorithm 2. $\mathcal{F}(V_s, F_s)$

Input: node mask V_s, feature mask F_s
Output: RDT-Fidelity for the given masks
1: **for** $i = 0, \ldots,$ samples **do**
2: Set $Y_{\{V_s,F_s\}}$, i.e., fix the selected values and otherwise retrieve random values from the respective columns of $\mathcal{X}$
3: **if** $\Phi(Y_{\{V_s,F_s\}})$ matches the original prediction of the model **then**
4: correct$+ = 1$
5: **return** $\frac{\text{correct}}{\text{samples}}$

Theorem 2. *ZORRO has the following properties.*

1) *ZORRO retrieves explanation with at least RDT-Fidelity* τ

$$\mathcal{F}(V_s, F_s) \geq \tau.$$

2) *The runtime of ZORRO is independent of the size of the graph. The runtime complexity of ZORRO for retrieving an explanation is given by*

$$O(t \cdot \max(|V_n|, |F|)),$$

where t *is the run time of the forward pass of the GNN* Φ.

3) *For any retrieved explanation* $\mathcal{S}$ *and* $\tau \geq 0.5$, *the stability score is* $\gamma(\mathcal{S}) \geq \frac{1}{1+\tau(1-\tau)}$.

For the proof and the discussion on the choice of noise distribution, we refer to Appendix C.2, available in the online supplemental material.

Discussion. We note that the explanations returned by ZORRO have *high validity*, *sparsity*, and *stability*. First, the RDT-Fidelity, which ZORRO tries to maximize is, by definition, the expected validity of the perturbed input. An increase in fidelity is, therefore, a result of an increase in validity at individual realizations of the perturbed input. Second, by Theorem 1 high fidelity explanations lead to a higher stability score.

Relation to Counterfactual Explanations. A few recent works focus on finding counterfactual explanations for the task of graph classification [2] and link prediction [16]. The goal is to find an explanation such that removing that explanation leads to a change in the model's decision. While it is symmetrical to our goal of finding an explanation that best preserves the prediction power of the model quantified via RDT-Fidelity, we are different from the above works. First, none of these works consider the node classification task for which an explanation needs to be generated corresponding to a query node. Therefore, it is not trivially clear if their strategies would also always lead to sparse counterfactual explanations for node classification. Second, we observe that for node classification there are in fact multiple explanations possible (see Appendix C.1), available in the online supplemental material. Such phenomena have also been observed for other data types, and models [5]. High fidelity explanations, as in our case, therefore, cannot be directly used as counterfactual explanations, at least for the task of node classification. One can construct a counterfactual by taking the union of multiple explanations.

6 Experimental Setup

The evaluation of post-hoc explanation techniques has always been tricky due to the lack of ground truth. Specifically, for a model prediction, collecting the ground truth explanation is akin to asking the trained model what it was thinking about – an impossibility and hence a dilemma. There is no clear solution to the ground-truth dilemma. However, previous research has attempted varying experimental regimes, each with its simplifying assumptions. We conduct a comprehensive set of experiments adopting the three dominant existing experimental regimes from the literature – *real-world graphs with unknown ground truth*, *remove and retrain*, *and synthetic graphs with known ground truth*. Later on, we will reflect on the limitations of their assumptions and the threats they might pose to our results' validity.

6.1 Evaluation Without Ground Truth

In the absence of ground truth for explanations, we can still evaluate posthoc explanations using the desirable properties of the explanations introduced by us, i.e., sparsity, stability (quantified via RDT-Fidelity), and validity:

RQ 1. How effective is ZORRO as compared to existing methods in terms of sparsity, RDT-Fidelity, and validity?

Note that these metrics are not always correlated. For example, an explanation can have a high validity score but low stability or RDT-Fidelity. In the following, we describe the real-world datasets that we use to compare explanations.

Datasets and GNN Models. We use the most commonly used datasets *Cora*, *CiteSeer* and *PubMed* [43]. We evaluate our approach on four different two-layer graph neural networks: *GCN* [20], graph attention network (*GAT*) [39], the approximation of personalized propagation of neural predictions (*APPNP*) [21], and graph isomorphism network (*GIN*) [42]. We evaluate these combinations with respect to *validity*, *sparsity*, and *RDT-Fidelity* for 300 randomly selected query nodes. To calculate node sparsity for those approaches which retrieve soft edge masks, such as GNNExplainer, we follow [30] and create node masks by distributing the edge mask value equally onto the endpoint of the respective edges. For example, if a particular edge (u, v) the corresponding edge mask has a value of 0.5, then nodes u and v would be given a node mask of 0.25 each.

In Appendix D, available in the online supplemental material, we provide additional experimental results in which we investigate: i) the effect of the number of samples used for calculating the RDT-Fidelity in ZORRO, ii) further variations of the RDT-Fidelity threshold, iii) explanations using four additional metrics proposed by [47], and iv) the impact of larger computational graphs on explanation approaches by using the Amazon Computers dataset [33].

6.2 Remove and Retrain

In this experimental regime, we follow the remove-and-retrain (or ROAR) paradigm of evaluating explanations [12] that is based on retraining a neural network based on the explanation outputs. ROAR removes the fraction of input features deemed to be the most important according to each explainer and measures the change to the model accuracy upon retraining. Thus, the most accurate explainer will identify inputs as necessary whose removal causes the most damage to model performance relative to all other explainers. Note that, unlike the other evaluation schemes, first, ROAR is a global approach in that it forces a fixed set of features to be removed. Second, ROAR involves retraining the model, whereas other approaches have interventions purely on the outputs of the trained model.

RQ 2. How effective is ZORRO when its output explanations are used for retraining a new GNN model?

6.3 Evaluation With Ground Truth

Although it is hard to obtain ground-truth data from real-world datasets, previous works have constructed synthetic datasets with known subgraph structures that GNN models learn to predict the output label [44]. We consider the only synthetic dataset proposed in [44] having features called BA-Community. First, we create a community with a base Barabási-Albert (BA) graph and attach a five-node house graph to randomly selected nodes of the base graph. Nodes are assigned to one of the eight classes based on their structural roles and community memberships. For example, there are three functions in a house-structured motif: the house's top, middle, and bottom nodes. Following [44], only the class assignments of the house nodes have to be explained, and the respective house is regarded as "explanation ground truth".

Node Features for Synthetic Graphs. Nodes have normally distributed feature vectors. Each node has eight feature values drawn from $\mathcal{N}(0,1)$ and two features drawn from $\mathcal{N}(-1,.5)$ for nodes of the first community or $\mathcal{N}(1,.5)$ otherwise. The feature values are normalized within each community, and within each community, 0.01 % of the edges are randomly perturbed. Note that for reproducibility, we strictly follow the published implementation of GNNExplainer. For the known ground-truth regime, we are interested in answering the following research question:

RQ 3. Are Zorro's explanations accurate, precise and faithful to available ground truth explanations?

6.3.1 Metrics

To compare against the known ground truth, we use various metrics proposed earlier in literature for synthetic datasets – *accuracy*, *precision*, *faithfulness*.

Accuracy measures the fraction of correctly classified nodes in the explanation. Note that only reporting accuracy as a metric does not portray the complete picture. For example, in our imbalanced dataset of five positive nodes (the house motif), out of 100 other nodes in the computational graph, high accuracy can be achieved by a trivial selection of five neighbors or sometimes even none. Therefore, we also report the precision value, which emphasizes the fraction of correct predictions:

Precision is defined as the fraction of returned nodes that are also in the explanation set. Precision provides more reliable results than accuracy when the input is much larger than the explanation. To compute accuracy and precision for baselines, we transform the baselines' results into a node mask of the five most important nodes, which is the size of the explanation ground truth.

Faithfulness. Faithfulness is based on the assumption that a more accurate GNN leads to more precise explanations [30]. We measure faithfulness by comparing the explanation performance of the GNN model at an intermediate training epoch against the fully trained GNN (final model trained until convergence). Specifically, we generate two ranked lists corresponding to test accuracy and precision of retrieved explanations at different epochs. We then compute faithfulness as the rank correlation of these two lists measured using Kendall's tau τ_{Kendall}.

6.4 Baselines and Competitors

For a comprehensive quantitative evaluation we chose our baselines from the three different categories of post-hoc explanations models consisting of (i) soft-masking approaches like *GNNExplainer*, which returns a continuous feature and edge mask and *PGE [24]* learns soft masks over edges in the graph (ii) surrogate model based hard-masking approach, *PGM* [40], which returns a binary node mask, iii) Shapely based hard masking approach SubgraphX [48], which returns a subgraph as an explanation, and (iv) gradient-based methods *Grad & GradInput* [34] which utilize gradients to compute feature attributions. Specifically, we take the gradient of the rows and columns of the input feature matrix X, which corresponds to the features' and nodes' importance. For GradInput, we also multiply the result element-wise with the input. In the case of PGM, we use the author's default settings to choose the best node mask. Besides, we employ an *empty explanation* as the naive baseline. We could only run SubgraphX for the small synthetic dataset, due to its long runtime (see Appendix B), available in the online supplemental material.

ZORRO *Variants.* For our approach ZORRO, we retrieved explanations for the thresholds $\tau = .85$ and $\tau = .98$ with $K = 10$. All RDT-Fidelity values were calculated based on 100 samples.

We refer to Appendix B, available in the online supplemental material, and the available implementation for further details of the models and the training of the GNNs.

7 EXPERIMENTAL RESULTS

In presenting our experimental results, we begin with RQ 1 that relates to the regime where we consider real-world datasets but without ground-truth explanations in Section 7.1. Continuing with the real-world datasets, we will

TABLE 1
Analysis of the Average *Sparsity (Definition 2)*, *RDT-Fidelity (Definition 4)*, and *Validity* (Definition 1) of the Explanations

Metric	Method	Cora				CiteSeer				PubMed			
		GCN	GAT	GIN	APPNP	GCN	GAT	GIN	APPNP	GCN	GAT	GIN	APPNP
Features-Sparsity	GNNExplainer	7.27	7.27	7.27	7.27	8.21	8.21	8.21	8.21	6.21	6.21	6.21	6.21
	Grad	4.08	4.22	4.45	4.08	4.19	4.28	4.41	4.18	4.41	4.51	4.89	4.46
	GradInput	4.07	4.25	4.37	4.08	4.17	4.29	4.33	4.17	4.41	4.51	4.92	4.47
	Zorro ($\tau = .85$)	**1.91**	**2.29**	**3.51**	**2.26**	**1.81**	**1.84**	**3.67**	**1.97**	**1.60**	**1.52**	**2.38**	**1.75**
	Zorro ($\tau = .98$)	2.69	3.07	4.34	3.18	2.58	2.60	4.68	2.78	2.55	2.58	3.21	2.86
Node-Sparsity	GNNExplainer	2.48	2.49	2.56	2.51	1.67	1.67	1.70	1.68	2.7	2.71	2.71	2.71
	PGM	2.06	1.82	**1.66**	1.99	1.47	1.59	**1.10**	1.54	1.64	1.16	**1.62**	2.93
	PGE	1.86	1.86	1.78	1.94	1.48	1.40	1.36	1.41	1.91	1.81	1.85	1.92
	Grad	2.48	2.34	2.25	2.35	1.70	1.61	1.55	1.60	2.91	2.76	3.11	2.73
	GradInput	2.53	2.43	2.23	2.41	1.61	1.58	1.54	1.52	3.02	2.94	3.41	2.81
	Zorro ($\tau = .85$)	**1.28**	**1.30**	1.90	**1.16**	**1.05**	**0.92**	1.36	**0.83**	**1.07**	**0.87**	1.77	**0.79**
	Zorro ($\tau = .98$)	1.58	1.59	2.17	1.48	1.26	1.09	1.58	1.07	1.51	1.31	2.18	1.25
RDT-Fidelity	GNNExplainer	0.71	0.66	0.52	0.65	0.68	0.69	0.51	0.62	0.67	0.73	0.67	0.72
	PGM	0.84	0.77	0.60	0.89	0.92	0.93	0.73	0.95	0.78	0.69	0.74	0.96
	PGE	0.50	0.53	0.35	0.49	0.64	0.60	0.51	0.61	0.49	0.61	0.56	0.50
	Grad	0.15	0.18	0.19	0.17	0.17	0.19	0.28	0.18	0.37	0.43	0.42	0.37
	GradInput	0.15	0.18	0.18	0.16	0.16	0.18	0.26	0.17	0.36	0.42	0.42	0.36
	Empty Explanation	0.15	0.18	0.18	0.16	0.16	0.18	0.26	0.17	0.36	0.42	0.42	0.36
	Zorro ($\tau = .85$)	0.87	0.88	0.86	0.88	0.87	0.86	0.87	0.86	0.86	0.88	0.88	0.87
	Zorro ($\tau = .98$)	**0.97**	**0.97**	**0.96**	**0.97**	**0.97**	**0.97**	**0.97**	**0.96**	**0.96**	**0.97**	**0.97**	**0.96**
Validity	GNNExplainer	0.89	0.95	0.83	0.84	0.87	0.92	0.58	0.93	0.60	0.81	0.71	0.87
	PGM	0.89	0.90	0.64	0.94	0.95	0.95	0.76	0.97	0.86	0.80	0.62	0.97
	PGE	0.51	0.54	0.34	0.45	0.62	0.59	0.54	0.62	0.51	0.61	0.57	0.48
	Grad	0.26	0.25	0.15	0.18	0.28	0.25	0.12	0.26	0.36	0.49	0.50	0.38
	GradInput	0.22	0.22	0.12	0.17	0.18	0.16	0.08	0.19	0.36	0.49	0.50	0.37
	Empty Explanation	0.22	0.22	0.11	0.17	0.18	0.16	0.08	0.19	0.36	0.49	0.50	0.37
	Zorro ($\tau = .85$)	**1.00**	**1.00**	0.83	**1.00**	**1.00**	**1.00**	0.77	**1.00**	0.90	**1.00**	0.84	**1.00**
	Zorro ($\tau = .98$)	**1.00**	**1.00**	**0.90**	**1.00**	**1.00**	**1.00**	**0.91**	**1.00**	**0.98**	**1.00**	**0.87**	**1.00**

The smaller the explanation size larger is the sparsity. As stability can be directly derived from RDT-Fidelity and increases with RDT-Fidelity > 0.5 *(see Theorem 1), it suffices to compare RDT-Fidelity to ensure stability. PGM and PGE are not included in the feature sparsity because they don't retrieve feature masks.*

discuss the global impact of explanation approaches when GNN models are retrained based on the explanations in Section 2. To our knowledge, we are the first to evaluate GNN-explanation approaches in the retraining setup. Finally, in Section 7.3, we will check the effectiveness of Zorro on synthetic datasets where ground-truth for explanations is known.

7.1 Evaluation With Real-World Data

To answer RQ 1, we evaluate Zorro's performance on three standard real-world datasets – *Cora*, *CiteSeer* and *PubMed*. As discussed in the last section, real-world datasets do not have accompanying ground-truth explanations. Instead, the results of our experiments are summarized in Table 1 where we compare the performance of various explanation methods in terms of validity, sparsity, and RDT-Fidelity.

Validity and RDT-Fidelity. We re-iterate that RDT-Fidelity measures the stability of the explanations. The validity, on the other hand, measures if the explanation alone retains the same class predictions. We first observe that gradient-based approaches obtain low RDT-Fidelity and validity compared to other soft-masking baselines like PGM and GNNExplainer. We observe that even the empty baseline achieves validity in the same range of $0.11 - 0.50$ as gradient-based methods. Hence, selecting no nodes and no features in the explanation, as done in the empty explanation baseline, yields similar performance as the gradient-based explanations. This result also establishes the superiority of GNN-specific explanation methods as PGM and GNNExplainer. Interestingly, while GNNExplainer outperforms PGM for Cora in terms of validity, PGM finds overall more stable explanations and shows higher RDT-Fidelity. On the other side, PGE performs worst out of all masks-based explainers.

Since Zorro optimizes for RDT-Fidelity, we expectantly deliver high performance for RDT-Fidelity. However, Zorro also convincingly outperforms all the existing baseline approaches for validity even if it is not explicitly optimized for validity. Additionally, a significant result here is that our heuristic yet efficient greedy procedure is already sufficient to produce near-optimal validity and RDT-Fidelity values.

Node and Feature Sparsity. Note that we differentiate between node and feature sparsity because explanation methods like PGM do not produce feature attributions. Moreover, we report the sparsity as the effective explanation size that is the entropy of the retrieved masks. The larger the explanation size, the lower will be the sparsity. First, we compare soft-masking approaches, i.e., gradient-based approaches, PGE, and GNNExplainer. We observe that the feature sparsity of GNNExplainer, somewhat surprisingly, is less sparse than even gradient-based approaches. Since PGE and GNNExplainer return soft edge

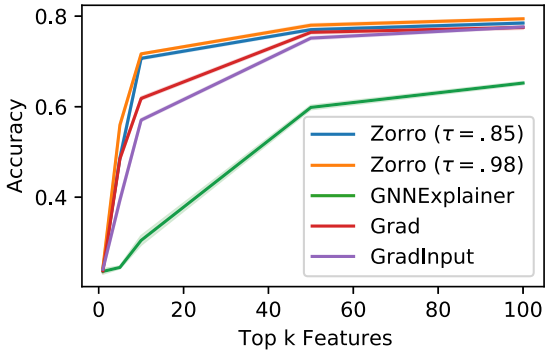

Fig. 3. Test accuracy after retraining GCN on Cora based on the top k features. We repeated the retraining 20 times, report the mean, and observed a variation of below .001.

masks, we compute the corresponding node mask as a sum of the masks of the edges which contain the corresponding node. In terms of node sparsity, PGE outperforms all other soft-mask-based approaches. As PGE does not produce a feature mask, in other words, it selects all features, feature sparsity is not provided. A low sparsity for soft-masking approaches implies a near-uniform feature attribution and consequently lower interpretability. On the other hand, explanations produced by ZORRO and PGM are hard masks. Since PGM only retrieves node masks, only a comparison based on node sparsity is possible. PGM outperforms with respect to node sparsity all soft masking approaches. However, for all cases, but GIN, ZORRO retrieves even sparser node masks.

We see that ZORRO produces significantly sparser explanations in comparison to soft-masking approaches. Between the variants of ZORRO, the explanations of ZORRO for $\tau = 0.85$ are expectedly lower in sparsity than for $\tau = 0.98$ that is a more constrained version of ZORRO. However, note that a lower sparsity comes with an advantage of higher RDT-Fidelity and validity.

Key Takeaways. Our crucial takeaway from this experiment is that ZORRO convincingly outperforms all other explanation methods across all datasets and GNN models. To answer RQ 1 quantitatively, we report the average improvement of ZORRO ($\tau = 0.98$) (with respect to the best performing baseline for each metric, model, and dataset): ZORRO achieves a reduction of 72% and 94% in the effective node, respectively, feature explanation size; increase by 20% in RDT-Fidelity and 11% in validity of explanations.

7.2 Evaluation With Remove and Retraining

We now present the results that estimate the global relevance of explanations by adapting the ROAR technique as already described earlier in Section 6.1. In our setup, given (a) a training set, (b) a GNN model that we want to explain and, (c) an explanation method, we retrieved explanations for each node in the training set. Next, we sum all feature masks corresponding to the retrieved explanations (of all training nodes) and choose the top-k features based on the aggregated value. For hard masks, this procedure is equivalent to selecting the top-k most frequently retrieved features. Finally, we retrain the GNN model again on the same training set but with the selected top-k features. Fig. 3 reports the performance drop in the model's test accuracy after the remove-and-retrain procedure. Note that the fewer features are needed to achieve similar test accuracy, the better the explanations' quality.

TABLE 2
Performance of the Node Explanation on the Synthetic Dataset

Method	Prec. ↑	Acc. ↑	Sparsity ↓	RDT-Fidelity ↑
GNNExplainer	0.40	0.81	1.68	0.63
PGM	0.75	0.93	2.30	0.81
PGE	0.19	0.21	1.73	0.58
SubgraphX	0.72	0.94	1.25	0.82
Grad	0.87	0.95	1.61	0.70
GradInput	0.89	**0.96**	1.61	0.56
$\emptyset$ - Explanation	0.00	0.84	0.00	0.55
ZORRO ($\tau = .85$)	**0.95**	0.90	**0.65**	0.91
ZORRO ($\tau = .98$)	0.90	0.90	1.04	**0.98**

The sparsity is calculated for the retrieved node mask. The high accuracy with empty explanation by large size of negative set. This also points to the pitfall of using Accuracy alone as the measure of evaluating explanations when ground truth is available.

We report our results for $k \in \{1, 5, 10, 50, 100\}$ most essential features using GCN as the GNN and over the CORA dataset. First, we observe that using only the top-10 important features using ZORRO ($\tau = .98$) already achieves a test accuracy of 0.72 compared to 0.79 on all 1433 features. Selecting 100 features however using ZORRO ($\tau = .98$), causes only a minor performance drop of $\Delta < 0.01$. Similar to Table 1, Grad achieves slightly better results than GradInput. Interestingly, GNNExplainer performs poorly, and the possible reason for this is its non-sparse feature masks (as seen in the previous section). Since PGM does not retrieve feature masks, it could not be evaluated in this setting.

To answer RQ 2, we find that ZORRO effectively chooses good global features, aggregated from ZORRO's local explanation, in comparison to other explanation approaches. Surprisingly, the gradient-based methods outperform GNN-specific GNNExplainer approaches. This is possible because we are experimenting with feature masks (and not node masks) and that gradient-based approaches are optimized for non-relational models.

7.3 Evaluation With Ground-Truth

For synthetic datasets, unlike real-world datasets, we have the liberty of having known ground truth explanations (GTE). We report *accuracy* and *precision* of explanations by comparing them against the GTE in Table 2. Note that for soft masking approaches, like GNNExplainer, hard masks need to be constructed by a discretizing step that is choosing top-k important attributions. In our experiments, we strengthen the soft-masking baselines by setting the k to the exact size of the ground truth. In addition, we also report the sparsity and RDT-Fidelity of corresponding explanations.

We observe that while the gradient-based methods achieve the highest accuracy, ZORRO achieves the best precision, sparsity, and RDT-Fidelity. The decrease in accuracy is due to two reasons. First, the higher accuracy of gradient-based methods is due to our decision to discretize the soft-masking baselines by allowing them active knowledge of the GTE size. On the other hand, ZORRO, natively outputs hard masks agnostic to the GTE size. PGE performs worst

TABLE 3
Experiments on Faithfulness According to [30] Measured With Kendall's tau τ_{Kendall} of the Retrieved Explanation Precision and Test Accuracy

Method	1	200	400	600	1400	2000	τ_{Kendall}
GNNExplainer	0.50	0.54	0.41	0.40	0.37	0.40	−0.73
PGM	0.83	0.47	0.68	0.71	0.76	0.75	0.20
PGE	0.20	0.19	0.23	0.21	0.23	0.20	0.36
Grad	**0.94**	0.80	0.62	0.73	0.84	0.87	0.07
GradInput	0.88	0.89	0.78	0.79	0.87	0.89	0.07
ZORRO ($\tau = .85$)	0.00	**0.92**	**0.88**	**0.93**	**0.94**	**0.94**	**0.73**
ZORRO ($\tau = .98$)	0.00	0.90	0.85	0.84	0.87	0.90	0.47

To simulate different model performances, we saved the GCN model during different epochs on the synthetic dataset. For ZORRO $\tau = .85$*, the ordering of the explanations' performances nearly perfectly align with the order of the models performance.*

in terms of precision and accuracy. SubgraphX retrieves explanations similar to Grad, but has the drawback of very high runtime, see Appendix B, available in the online supplemental material. To fully answer RQ 3, we also compared the achieved faithfulness of the explanation methods. Table 3 shows the model's and explainers' accuracies at different epochs. Even at epoch 1 with random weights, we see that the baselines achieve high precision on this synthetic dataset.

ZORRO achieves the first accuracy peak at 200 epochs, where the model still cannot differentiate the motif from the BA nodes. A similar peak at epoch 200 is observed for the GNNExplainer and GradInput. Moreover, the gradient-based methods achieve the best or close to the best performance for the untrained GCN. From these observations, we conclude that the underlying assumption – the better the GNN, the better the explanation – not necessarily need to hold. In this synthetic setting, the GNN only needs to differentiate the two communities, and all but one explanation method can explain the house motif.

In conclusion, for RQ 3, ZORRO outperforms all baselines in terms of precision and faithfulness while gradient-based methods achieve the highest accuracy. Note that the good performance of gradient-based approaches conflicts with our conclusions when experimenting with real-world data. We believe that this is a possible threat to be aware of while evaluating explanations. Specifically, some explanation methods perform admirably in more straightforward and synthetic cases but are not robust and do not generalize well when used in real-world scenarios. However, ZORRO is quite robust to different types of models, data, and evaluation regimes.

8 UTILITY OF EXPLANATIONS

One of the motivations of post-hoc explanations is to derive insights into a model's inner workings. Specifically, we use explanations to analyze the behavior of different GNNs with respect to *homophily*. GNNs are known to exploit the homophily in the neighborhood to learn powerful function approximators. We use the retrieved explanations by ZORRO to verify the models' tendency to use homophily for node classification and identify the model's mistakes.

Formally, we define the *homophily* of the node as the fraction of the number of its neighbors, which share the *same label* as the node itself. In what follows, we use homophily to refer to the homophily of a node with respect to the selected nodes in its explanation. *True homophily* is computed based on the true labels of the neighbor nodes. Similarly, *predicted homophily* is computed based on the predicted labels of the neighboring nodes.

8.1 Wrong Predictions Despite High Homophily

We start to investigate the joint density of true and predicted homophily of a given node. In Fig. 4, we illustrate the effect of connectivity and neighbors' labels on the model's decision for a query node (for the PubMed dataset). Several vertices corresponding to blue regions spread over the bottom of the plots have low predicted homophily. These nodes are incorrectly predicted, and their label differs from those predicted for the nodes in their explanation set. The surprising fact is that even though some of them have high true homophily close to 1, their predicted homophily is low. This also points to the usefulness of our found explanation in which we conclude that nodes influencing the current node do not share its label. So despite the consensus of GNNs reliance on homophily, they can still make mistakes for high homophily nodes, for example, when information from features is misaligned (or leads to a different decision) with that from the structure.

8.2 Incorrect High Homophily Predictions

We also note that for GIN and APPNP, we have some nodes with true homophily and predicted homophily close to 1 but are incorrectly predicted. This implies that the node itself and the most influential nodes from its computational graph have been assigned the same label. We can conclude

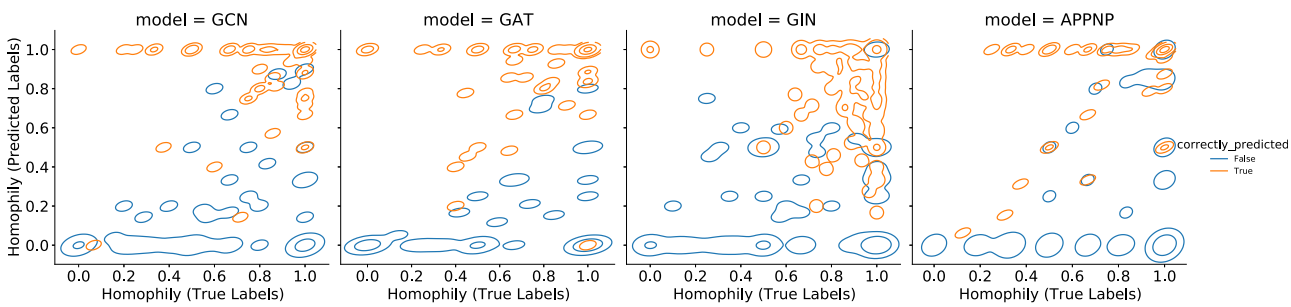

Fig. 4. Dataset - PubMed. The joint distribution of the homophily with respect to the nodes selected in the ZORRO's explanation ($\tau = .85$) with true and predicted labels. The orange contour lines correspond to the distributions for correctly predicted nodes, and the blue one corresponds to incorrectly predicted nodes.

that the model based its decision on the right set of nodes but assigned the wrong class to the whole group.

8.3 Influence of Nodes With Low Homophily

Nodes in the orange regions on the extreme left side of the plots exhibited low true homophily but high predicted homophily. The class labels for such nodes are correctly predicted. However, the corresponding nodes in the explanation were assigned the wrong labels (if they were assigned the same labels as that of the particular node in question, its predicted homophily would have been increased). The density of such regions in APPNP is lower than in GCN, implying that APPNP makes fewer mistakes in assigning labels to neighbors of low homophily nodes. For example, there are no nodes with true homophily 0, which incorrectly influenced its neighbors. These nodes can be further studied with respect to their degree and features.

9 CONCLUSION

We formulated the key properties a GNN explanation should follow: *validity*, *sparsity*, and *stability*. While none of these measures alone suffice to evaluate a GNN explanation, we introduce a new metric called RDT-Fidelity that along with a sparsity constraint reflects these desiderata into a single measure. We provide theoretical foundations of RDT-Fidelity from the area of *rate-distortion theory*. Furthermore, we proposed a simple combinatorial procedure ZORRO, which retrieves sparse *binary masks* for the features and relevant nodes while trying to optimize for fidelity. Our experimental results on synthetic and real-world datasets show massive improvements not only for fidelity but also concerning evaluation measures employed by previous works.

REFERENCES

[1] J. Baan, M. ter Hoeve, M. van der Wees, A. Schuth, and M. de Rijke, "Do transformer attention heads provide transparency in abstractive summarization?," 2019, *arXiv:1907.00570*.

[2] M. Bajaj et al., "Robust counterfactual explanations on graph neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 5644–5655.

[3] A. Binder, G. Montavon, S. Lapuschkin, K.-R. Müller, and W. Samek, "Layer-wise relevance propagation for neural networks with local renormalization layers," in *Proc. Int. Conf. Artif. Neural Netw.*, 2016, pp. 63–71.

[4] O.-M. Camburu, E. Giunchiglia, J. Foerster, T. Lukasiewicz, and P. Blunsom, "The struggles of feature-based explanations: Shapley values vs. minimal sufficient subsets," 2020, *arXiv:2009.11023*.

[5] B. Carter, J. Mueller, S. Jain, and D. Gifford, "What made you do this? Understanding black-box decisions with sufficient input subsets," in *Proc. 22nd Int. Conf. Artif. Intell. Statist.*, 2019, pp. 567–576.

[6] B. Chen, J. Hong, and Y. Wang, "The minimum feature subset selection problem," *J. Comput. Sci. Technol.*, vol. 12, no. 2, pp. 145–153, 1997.

[7] J. Chen, L. Song, M. J. Wainwright, and M. I. Jordan, "Learning to explain: An information-theoretic perspective on model interpretation," 2018, *arXiv:1802.07814*.

[8] A. Duval and F. D. Malliaros, "GraphSVX: Shapley value explanations for graph neural networks," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discov. Databases*, 2021, pp. 302–318.

[9] M. Fey and J. E. Lenssen, "Fast graph representation learning with PyTorch Geometric," in *Proc. Int. Conf. Learn. Representations Workshop*, 2019.

[10] T. Funke, T. Guo, A. Lancic, and N. Antulov-Fantulin, "Low-dimensional statistical manifold embedding of directed graphs," in *Proc. Int. Conf. Learn. Representations*, 2020.

[11] Y. Gao, T. Sun, R. Bhatt, D. Yu, S. Hong, and L. Zhao, "GNES: Learning to explain graph neural networks," in *Proc. IEEE Int. Conf. Data Mining*, 2021, pp. 131–140.

[12] S. Hooker, D. Erhan, P.-J. Kindermans, and B. Kim, "A benchmark for interpretability methods in deep neural networks," in *Proc. 33rd Int. Conf. Neural Inf. Process. Syst.*, 2019, Art. no. 873.

[13] Q. Huang, M. Yamada, Y. Tian, D. Singh, D. Yin, and Y. Chang, "GraphLIME: Local interpretable model explanations for graph neural networks," 2020, *arXiv:2001.06216*.

[14] M. Idahl, M. Khosla, and A. Anand, "Finding interpretable concept spaces in node embeddings using knowledge bases," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discov. Databases*, 2019, pp. 229–240.

[15] A. Jacovi and Y. Goldberg, "Aligning faithful interpretations with their social attribution," 2020, *arXiv:2006.01067*.

[16] B. Kang, J. Lijffijt, and T. De Bie, "ExplaiNE: An approach for explaining network embedding-based link predictions," 2019, *arXiv:1904.12694*.

[17] T. Kasanishi, X. Wang, and T. Yamasaki, "Edge-level explanations for graph neural networks by extending explainability methods for convolutional neural networks," in *Proc. IEEE Int. Symp. Multimedia*, 2021, pp. 249–252.

[18] M. Khosla, V. Setty, and A. Anand, "A comparative study for unsupervised network representation learning," *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 5, pp. 1807–1818, May 2021.

[19] P.-J. Kindermans et al., "The (un) reliability of saliency methods," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, Berlin, Germany: Springer, 2019.

[20] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. Int. Conf. Learn. Representations*, 2017.

[21] J. Klicpera, A. Bojchevski, and S. Günnemann, "Predict then propagate: Graph neural networks meet personalized PageRank," in *Proc. Int. Conf. Learn. Representations*, 2019.

[22] W. Lin, H. Lan, and B. Li, "Generative causal explanations for graph neural networks," in *Proc. 38th Int. Conf. Mach. Learn.*, 2021, pp. 6666–6679.

[23] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 4768–4777.

[24] D. Luo et al., "Parameterized explainer for graph neural network," 2020, *arXiv:2011.04573*.

[25] B. Perozzi, R. Al-Rfou, and S. Skiena, "DeepWalk: Online learning of social representations," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2014, pp. 701–710.

[26] P. E. Pope, S. Kolouri, M. Rostami, C. E. Martin, and H. Hoffmann, "Explainability methods for graph convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 10764–10773.

[27] D. Pruthi, M. Gupta, B. Dhingra, G. Neubig, and Z. C. Lipton, "Learning to deceive with attention-based explanations," 2019, *arXiv:1909.07913*.

[28] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?" explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2016, pp. 1135–1144.

[29] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-precision model-agnostic explanations," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, Art. no. 187.

[30] B. Sanchez-Lengeling et al., "Evaluating attribution for graph neural networks," in *Proc. 34th Int. Conf. Neural Inf. Process. Syst.*, 2020, Art. no. 495.

[31] M. S. Schlichtkrull, N. De Cao, and I. Titov, "Interpreting graph neural networks for NLP with differentiable edge masking," 2020, *arXiv:2010.00577*.

[32] T. Schnake et al., "Higher-order explanations of graph neural networks via relevant walks," 2020, *arXiv:2006.03589*.

[33] O. Shchur, M. Mumme, A. Bojchevski, and S. Günnemann, "Pitfalls of graph neural network evaluation," 2018, *arXiv:1811.05868*.

[34] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 3145–3153.

[35] C. R. Sims, "Rate–distortion theory and human perception," *Cognition*, vol. 152, pp. 181–198, 2016.

[36] J. Singh, M. Khosla, W. Zhenye, and A. Anand, "Extracting per query valid explanations for blackbox learning-to-rank models," in *Proc. ACM SIGIR Int. Conf. Theory Inf. Retrieval*, 2021, pp. 203–210.
[37] J. Singh, Z. Wang, M. Khosla, and A. Anand, "Valid explanations for learning to rank models," 2020, *arXiv:2004.13972*.
[38] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 3319–3328.
[39] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *Proc. Int. Conf. Learn. Representations*, 2018.
[40] M. N. Vu and M. T. Thai, "PGM-explainer: Probabilistic graphical model explanations for graph neural networks," in *Proc. 34th Int. Conf. Neural Inf. Process. Syst.*, 2020, Art. no. 1025.
[41] L. Wolf, T. Galanti, and T. Hazan, "A formal approach to explainability," in *Proc. AAAI/ACM Conf. AI Ethics Soc.*, 2019, pp. 255–261.
[42] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, "How powerful are graph neural networks?," in *Proc. Int. Conf. Learn. Representations*, 2019.
[43] Z. Yang, W. W. Cohen, and R. Salakhutdinov, "Revisiting semi-supervised learning with graph embeddings," 2016, *arXiv:1603.08861*.
[44] R. Ying, D. Bourgeois, J. You, M. Zitnik, and J. Leskovec, "GNN explainer: A tool for post-hoc explanation of graph neural networks," 2019, *arXiv:1903.03894*.
[45] J. Yoon, J. Jordon, and M. van der Schaar, "INVASE: Instance-wise variable selection using neural networks," in *Proc. Int. Conf. Learn. Representations*, 2018.
[46] H. Yuan, J. Tang, X. Hu, and S. Ji, "XGNN: Towards model-level explanations of graph neural networks," in *Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2020, pp. 430–438.
[47] H. Yuan, H. Yu, S. Gui, and S. Ji, "Explainability in graph neural networks: A taxonomic survey," 2020, *arXiv:2012.15445*.
[48] H. Yuan, H. Yu, J. Wang, K. Li, and S. Ji, "On explainability of graph neural networks via subgraph explorations," 2021, *arXiv:2102.05152*.
[49] Z. Zhang, J. Singh, U. Gadiraju, and A. Anand, "Dissonance between human and machine understanding," *Proc. ACM Human-Comput. Interaction*, vol. 3, 2019, Art. no. 56.

Thorben Funke received the PhD degree from the University of Bremen, Germany. Currently, he is a post doctoral researcher with L3S Research Center. His main research interests are machine learning algorithms on graphs, interpretable machine learning, and spatio-temporal models.

Megha Khosla is an assistant professor with Intelligent Systems Department, TU Delft, Netherlands. Her main research area is machine learning on graphs with focus on three key aspects of effectiveness, interpretability and privacy-preserving learning.

Mandeep Rathee received the master's degree in mathematics and computing from the Indian Institute of Technology, Patna, India. Currently, he is working toward the PhD degree with L3S Research Center, Leibniz University, Hannover. His main research interests include on interpretability in graph neural networks.

Avishek Anand is an associate professor with the Web Information Systems (WIS), Software Technology (ST) Department, Delft University of Technology (TU Delft). He is also a member of the L3S Research Center, Hannover, Germany. One of his main research focus is interpretability of machine learning models with focus on representations from discrete input like text and graphs.