There are many contrastive (opposite) categories of explanations, e.g.,

* + **Global vs. Local (Instance-level vs. Class-level)**
  + **Instance-level vs. Model-level**
  + Factual vs. CounterFactual
  + Output of explainability method (node, node features, edge, subgraph)
  + Self-explainable vs. Post-hoc

GNN explainability methods are generally proposed considering one of the two contrastive classes in mind (e.g., a method is proposed only as instance-level, and not model-level).

Certain metrics are also only suitable or proposed considering one of the two contrastive classes in consideration (Examples of such metrics?).

So, we would like to investigate the following:

* metric adaptations, some metrics are specific for one of the contrastive classes (examples of such metrics?), whether we can use or adapt it for both classes?
* method adaptations, some methods are specific for one of the contrastive classes, whether we can use or adapt it for both classes? e.g., an instance-level method’s adaptability to model-level.
* if a method performs well on a specific class, how it performs for the other contrastive category? Maybe this was never explored in the existing literature?
* efficiency and scalability of contrastive classes (general trends, e.g., are instance-level methods generally faster compared to model-level methods?)

**Feb 13 meeting notes and discussions:**

**Global vs. Local**

**Formal Definitions**

Global:

Local:

Tingyang: Let $X = \{x\_1, x\_2, ..., x\_m\}$ be a set of graph samples, where each sample $x\_i$ belongs to a class $C\_k$ within a finite set of classes $C = \{C\_1, C\_2, ..., C\_k\}$. For each sample $x\_i$, an explanation $E(x\_i)$ is generated such that $E(x\_i)$ is independent of $E(x\_j)$ for all $i != j$, irrespective of whether $x\_i$ and $x\_j$ belong to the same class $C\_k$ there will be $|C\_k|$ explanations generated. These explanations, $\{E(x\_{k1}), E(x\_{k2}), ..., E(x\_{k|C\_k|})\}$, for the samples in class $C\_k$, reflecting the principle that the generation of explanations is sample-based and strictly independent across individual samples, without regard to their class membership.

Ehsan: In global explanation methods, where graph patterns are generated to maximize the predicted probability for a certain class and use such graph patterns to explain the entire class [GNES, ICDM 2021]. For a given GNN $\Phi$ and target class y, global explainers target to produce an explanation for $\Phi$ such that explains behavior of $\Phi$ for class y [Global Concept-Based, AAAI 2023].

Let $X = \{x\_1, x\_2, ..., x\_m\}$ be a set of graph samples, where all samples belong to the same class $C\_k$ within a finite set of classes $C = \{C\_1, C\_2, ..., C\_k\}$. For each class $C\_i$, an explanation $E(C\_i)$ is generated such that $E(C\_i)$ is independent of $E(C\_j)$ for all $|k-1|$ classes. Thus, there will be $|k|$ explanations generated. These explanations, $\{E(C\_1), E(C\_2), ..., E(C\_k)\}$, for the classes in $X$, reflecting the principle that the generation of explanations is class-based and strictly independent across classes and dependent across individual samples from the same class.

**Potential Metrics**

Fidelity+, Fidelity-, Sparsity, Contrastivity, Stability, Consistency, Accuracy, Running Time

1. Local to Global: Mean and Std of Evaluation metrics for a class

2. Global to Local: Projection of global explanations to local scale [XGNN, a global method, does not work in projection-based manner] - global instance level is fine

What about global model level?

*AK: (1) Still unclear what are the metrics used in global explainability papers? They measure quality directly, and not by projection, right?*

*Tingyang: There are some related things, such as:*

*1* ***predicted Class Probability****: represents the probability that the explanation generated by the global method for a class is classified by GNN as that class.*

*2 Some metrics measure* ***discriminability*** *between different classes (Data-awareness, coverage, Concept Purity, Contrastivity)****.***

*3* ***utility:*** *use the explanations to retrain a GNN model from scratch and report its performance on the test set. It is an interesting metric that considers the performance of global methods from the perspective of model training accuracy.*

Ehsan:

**Rules:**

* There is no GLOBAL evaluation metric.
* We focus on NODEs and EDGES, not node features (for explanation generation: attribution scores).
* Applicability of an evaluation metric is dependent on:
  + How an explainer is trained (Local or Global)
  + How it will be tested (Globally or Locally or Both)
* Globally trained Instance-level methods can be tested locally (local explanation generation for test set).
* Globally trained Model-level methods that have generation-based nature, can not generate local explanations for test samples (e.g., XGNN); Some adaptations might be possible, but I don't have any concrete solution, yet.

**Evaluations for Global Methods**:

**Global Methods:**

* **Instance-Level:**
  + **PGExplainer:** Explanation AUC and Inference time on Mutagenicity and BA-2motifs dataset. (All based on explanation ground truth).
  + **GraphMask:**Precision, Recall, and F1. Referenced to (Li's faithfulness gold standard on the toy task, 2016).
  + **GCFExplainer:**
    - **Coverage:** The proportion of input graphs that have close counterfactuals from C (explanation for the class) under a given distance threshold.
    - **Cost:** The distance between the input graph and its counterfactual across the input graphs.
    - **Interpretability:** We quantify the interpretability as the size of recourse representation. (explanation size)
  + **Glocal:** 
    - **Faithfulness:**
      * **Fidelity:** By averaging the scores among instances in the respective functioning. Ratio of nodes in explanation to input graph size producted on the fid score.
      * **InFidelity:** By averaging the scores among instances in the respective functioning. Ratio of nodes in explanation to input graph size producted on the fid score.
  + **DAG:**
    - **Overall recognizability:** The GNN score of the target class is used as an evaluation metric in many existing model-level methods.
    - **Data-awareness:** measure the discrimination level for explanations by a pair of metrics (claimed to use VF2 to measure data-awareness on XGNN):
      * Faithfulness:
        + support
        + Denial
  + **GLGExplainer:** 
    - **To show robustness, three metrics are used (mean and std Table 2):**
      * **Fidelity:** Represents the accuracy of E-LEN in matching the predictions of the model f to explain.
      * **Accuracy:** represents the accuracy of the formulas in matching the ground-truth labels of the graphs.
      * **Concept Purity:** which is computed for every cluster independently and measures how good the embedding is at clustering the local explanations.
  + **GDM**:
    - **Utility**: Utility aims to verify whether generated interpretations can be utilized and leads to a well-trained GNN. Desired interpretations should capture the dominating patterns that guide the training procedure. For this protocol, we use the interpretive graphs to train a GNN model from scratch and report its performance on the test set. We call this accuracy utility.
    - **Fidelity**
* **Model-Level:**
  + **XGNN:** no baseline for quantitative comparisons and evaluations, qualitative evaluations on real-world and synthetic datasets. (size is a hyper-parameter)
  + **GNNInterpreter:**
    - **(1-Predicted Class Probability) by GNN:** mean and std. (Mean of Fidelity-) (size?)
    - **Training Time per Class**
  + **Global Concept-Based:**
    - **Neuron Importance**
    - **Neuron Correctness**
* **Meaning of these metrics? Applicable to other two categories above? (size?)**

*(2) Can contrastivity be defined at a global level directly?*

*Tingyang: Contrastivity was designed to capture the intuition that class-specific features highlighted by an explanation method should differ between classes. This is very natural and can be used for global methods without modification.*

Ehsan: Yes, only depends on how the explainer is trained and tested.

*(3) We did not discuss Stability, Consistency, Accuracy. What about them at both local and global level?*

**Stability**: An attribution method should be invariant to small changes in input features that do not affect an example’s class label or the model’s prediction. To assess stability, we make small graph perturbations on test set examples that leave the ground-truth attribution and predicted class label unchanged, and assess the degree of change in attribution accuracy (Sanchez’s benchmarking, NeurIPS 2020, this paper is the reference for Stability).

Given an input graph Gi, its explanations mi is regarded as the ground truth. Then, the input graph Gi is perturbed by small changes, such as attaching new nodes/edges, to obtain a new graph Gˆi. Note that Gi and Gˆi are required to have the same predictions. Then, the explanations of Gˆi is obtained, denoted as mˆi. By comparing the difference between mi and mˆi, we can compute the Stability score (Yuan’s taxonomy).

Stability is local by definition since it is dependent on the explanations of individual samples.

**Consistency**: The accuracy of an attribution technique should be consistent across high-performing model architectures. To test attribution consistency, we quantify the variability in attribution accuracy using the top 10% of models through a hyperparameter scan over model architectures (Sanchez’s benchmarking, NeurIPS 2020, this paper is the reference for Consistency).

Consistency does not seem to be an appropriate option for our experiments. It deals with hyperparameters of GNNs and measures how accuracy varies across different hyperparameters of a model.

???

**Potential Methods**

Global Methods: PGExplainer, GraphMask, XGNN, …

Local Methods: GNNExplainer, SubgraphX, PGMExplainer, …

Justification of selection? - total 9 methods (3 from each of local, global model, global instance)

Third-level model or basic comparisons like how confidently one wins, or on how many metrics one method wins competitions.

**Q1. Do adaptations of evaluation metrics and methods for these two contrastive categories generate high-quality results for the opposite category for which a method was not originally proposed?**

**Ehsan:**

Adaptations do not happen at the metrics level, they happen on the explanation stage. All evaluation metrics are computable iff the importance scores are node based for each graph (local and dependent on nodes), which is possible if we adapt all to local explanations on test dataset. Then, average and standard deviations of the evaluations on local scale can result in global values.

**Q2. efficiency and scalability of contrastive classes?**

Efficiency and scalability of the contrastive categories should be checked by different datasets. Except for MUTAG, better we test on one more small dataset, like isCyclic by Yuan’s taxonomy paper. All our comparisons on efficiency and scalability should be done on a global scale not local scale.

**Experiments**

Fidelity+, Fidelity-, Sparsity, Contrastivity, Stability, Accuracy, Consistency, Running Time.

Recently, Warmsley has introduced a new metric which is a combination of Fid+ and Fid-, named Harmonic Fid (in case we look for a substitution for Consistency).

All explanations are adapted to the local level. Only generation-based methods (from global and model-level) are not possible to be evaluated by these metrics (I haven’t seen any adaptations for this end, I have some ideas like size limits for explanations).

—---------------

**Experimental Setup**

- So, what is the difference between two spreadsheets, same information in both of them, presented differently, right?

- The list of metrics looks fine. (Scenario 1 and 2 in Meeting Minutes and Notes folder)

- It also needs to be discussed (formally) how those metrics are computed and compared for local and instance level, global and instance-level, global and model level?

Why is such comparison fair and apple-to-apple comparison?

Can you compute these metrics for all three categories and 3 explainability methods under each category?

cv=vision

M: size is < 10K

B: Million Scale

**Datasets:**

**Small Graph: 0 - 1k, Medium Graph 1k - 100k, Big Graph: 100k~**

**Small Dataset: 0 - 10k Medium Dataset: 10k - 100k, Big Dataset: 100k~**

| **Datasets** | **D-G** | **#NodeFeat** | **#Graphs** | **#Nodes(avg)** | **#dEdges(avg.)** | **#Classes** | **Domain** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| MUTAG | S-S | 7 | 188 | 18 | 40 | 2 | molecule |
| ENZYMES | S-S | 3 | 600 | 32.6 | 62.3 | 6 | biology |
| NCI1 | S-S | 37 | 4110 | 30 | 65 | 2 | biology |
| Graph-SST5 | M-S | 768 | 11855 | 20 |  | 5 | vision |
| Synthetic | S-B | 5 | 100 | 1000000 | 4000000 | 2 | synthetic |

**More choices:**

| **Datasets** | **# Node Feat** | **# Graphs** | **# Nodes (avg)** | **# dEdges(avg.)** | **# Classes** | **Domain** |
| --- | --- | --- | --- | --- | --- | --- |
| IsACyclic |  |  | 28 | 60 | 2 |  |
| Highschool |  |  |  |  |  |  |
| ENZYMES |  | 600 | 32.6 | 62.3 | 6 | biology |
| Reddit-binary |  | 2000 | 429.6 | 995 | 2 | social networks |
| Graph-SST2 |  | 70042 | 10.199 |  | 2 | text |

**So please have a look and let us know your ideas.**

**We first work on Local perspective, and compute 6 evaluation metrics on this part.**

**Implementation of the 6 evaluation metrics are ready.**

Fid+, Fid-, Contrastivity, Sparsity, Stability, Time.

**Then, we will complete the global perspective’s associated values.**

Subgraph iSomorphism, for global perspective to be implemented.

**Evaluation Metrics:**

**Fidelity+**: Fidelity was calculated to capture the intuition that occlusion of salient features identified through explanations should decrease classification accuracy. More precisely, we define fidelity as the difference in accuracy obtained by occluding all nodes with saliency value greater than **0.01** (on a scale 0 to 1). We then averaged the fidelity scores across classes for each method. [Pope’s benchmarking paper]. A large fidelity score indicates stronger counterfactual characteristics [RCExplainer].

**Fidelity-**: also known as Inverse Fidelity, assumes that the removal of insignificant features should not change GNN performance very much at all since they have little contribution to the GNNs decision [Warmsley’s paper]. Fidelity- based on prediction accuracy is calculated as:

**Contrastivity**: means the ratio of the Hamming distance between binarized heat-maps for positive and negative classes. The underlying idea behind contrastivity is that the highlighted features by an explanation method should vary across classes [Li’s benchmarking paper].

**Sparsity**: measures the fraction of structures or features that are important by explanation methods [Li’s benchmarking paper]

where represents the total number of features such as nodes, nodes features, and edges in the original graph model; while is the important features found by the explainable methods and it is a subset of ; N is the total number of samples.

**Stability**: infinitesimally small perturbations to an instance (which do not affect its model prediction) should not change its explanation drastically.

where is the explanation for u′, D(·) computes distance between two explanations, and δ is an infinitesimally small constant.

In all metrics, we have a base graph, and generate explanations for that base graph w.r.t the two classes, in all these metrics we need size of the base graph (math formula). The base graph and its explanations for two classes are the inputs for the evaluation metrics. All those 5 metrics: Fid+, Fid-, Contrastivity, Sparsity, Stability are instance-level. There is a base graph (a test graph) and two explanations (binary classification), then these three are used as input for different formulas (metrics). In XGNN and GNNInterpreter we don’t have that base graph, instead we have the entire class (multiple graphs with different sizes), which makes it impossible to investigate with respect to the two classes how explanations differ. For this end, for each test graph find the common EDGES in between the graph and explanations of two classes (intersection of each test graph with explanations of the two classes) and use this idea for evaluation metrics (if nothing in common, then 0). (summary: subgraph isomorphism of explanation and test graphs.) So, each test graph is considered as a base graph (iteratively) and checks which EDGES are in common between the base graph and explanations of two classes.

**Apr 2:**

Why does GNNExplainer perform better on GCN+GAP and DGCNN than GIN and DIFFPOOL on the mutag dataset(Fid+)?

Why does SubgraphX have better results on GIN and DIFFPOOL compared to GCN+GAP and DGCNN(Fid+)?

Why does PGMExplaienr have better results on GIN and DIFFPOOL compared to GCN+GAP and DGCNN(Fid+)?

Why does CF2 have better results on GIN, GCN+GAP, and DGCNN compared to DIFFPOOL(Fid+)?

Why does PGMExplainer have better results on all the models (Contrastivity)? PGMExplainer extracts independent elements in an input graph for each class and computes the Chi-Square test; this independent elements extraction enables it to deal with low complexity and better efficiency on contrastive elements.