The 38th Annual AAAI Conference on Artificial Intelligence

# Self-Interpretable Graph Learning with Sufficient and Necessary Explanations

*Jiale Deng, Yanyan Shen*

**Shanghai Jiao Tong University**

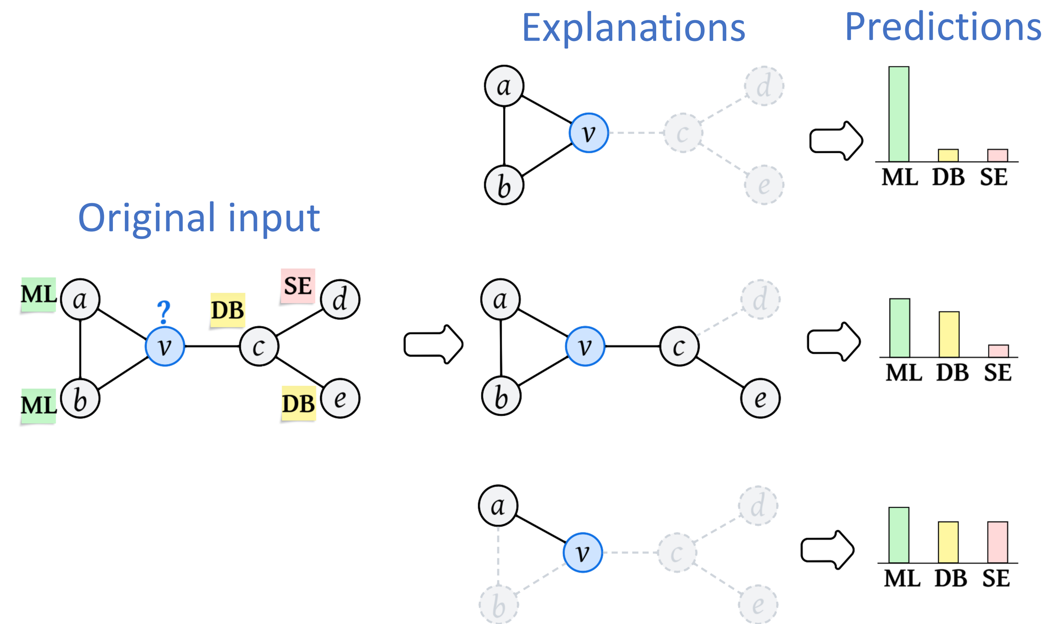
上海交通大學 SHANGHAI JIAO TONG UNIVERSITY

## 1. Background

- GNNs show strong performance in graph learning tasks but lack transparency.
- To **explain** the predictions of GNNs:
  - **Post-hoc methods** extract salient substructures from the input graph as explanations, but they can be biased and inconsistent.
  - **Self-interpretable methods** can provide unbiased explanations by generating built-in explanations and making predictions based on the explanations.

## 2. Motivation

- **Issues** of Self-Interpretable Methods:
  - The quality of explanations determines the model performance.
  - **An example** (predicting the research area of an author node in coauthor network):




**Sufficient & Necessary**: Ideal explanation capturing the rationale that *v, a, b form an ML research group (clique)*.

**Sufficient**: Contains the clique while *introduce noisy edges (e,c) and (c,v)*.

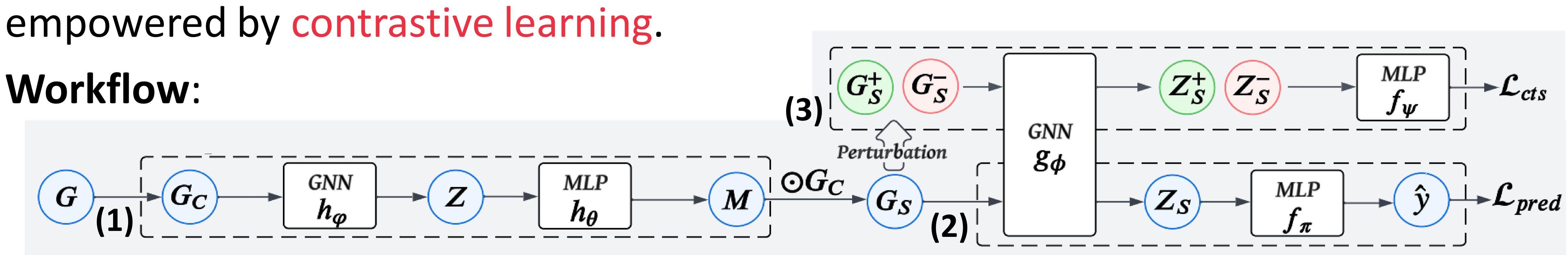
**Necessary**: *Missing the salient clique*.

***Compromised performance!***

- **Key Insight**: Promote the quality of explanations toward **both sufficiency and necessity directions**, encouraging the explanations to improve model performance.

## 3. Method

- Self-interpretable graph learning with SUfficient aNd NecessarY explanations (SUNNY-GNN) empowered by contrastive learning.
- **Workflow**:




(1) **Explanation generation**: inputs edge embeddings, outputs edge importance mask $M$.

(2) **Prediction** with generated explanation $G_S$.

(3) **Augmentations** on $G_S$:

- Positive samples $G_S^+$: Sample edges from $G_C \backslash G_S$ and add them to $G_S$.
- Negative samples $G_S^-$: Sample edges in $G_S$ and add them to $G_S$.
- **How to sample**? Simply sample by $M$ may lead to:
  - Trivial samples impair the self-supervision signals→Solution: Enhance contrastive signal.
    - **Distance coefficients** $\delta$: *perturbations on edges closer to the target node v tend to have a greater impact to v than those on farther edges.*
  - Unreliable samples mislead GNN training→Solution: Filter out unreliable ones with labels.
    - **Confidence coefficients** $\eta$: introducing noisy edges and removing irrelevant edges forms *untrustworthy* positive and negative samples, respectively.

- **Optimization**: $\min_{\Theta} \mathcal{L}_{pred} + \gamma \mathcal{L}_{cts}$ where ① Prediction loss: $\mathcal{L}_{pred} = -\frac{1}{|\mathcal{V}_{train}|} \sum_{v \in \mathcal{V}_{train}} \sum_{t=1}^{\mathcal{T}} \mathcal{Y}_{vt} \log \hat{\mathcal{Y}}_{vt}$

and ② Contrastive loss: $\mathcal{L}_{cts}(v) = \mathbb{E}\left[-\log \frac{\eta_i^+ \exp(z_S^\top z_{S_i}^+/\tau)}{\sum_j \eta_j^+ \exp(z_S^\top z_{S_j}^+/\tau) + \sum_k \eta_k^- \exp(z_S^\top z_{S_k}^-/\tau)}\right] \Rightarrow \mathcal{L}_{cts} = \frac{1}{|\mathcal{V}_{train}|} \sum_{v \in \mathcal{V}_{train}} \mathcal{L}_{cts}(v)$

- **Minimizing $\mathcal{L}_{cts}$ means**:
  - Pulling $G_S^+$ closer to $G_S$ → *turning $G_S$ to be more sufficient*.
  - Pushing $G_S^-$ distant from $G_S$ → *turning $G_S$ to be more necessary*.

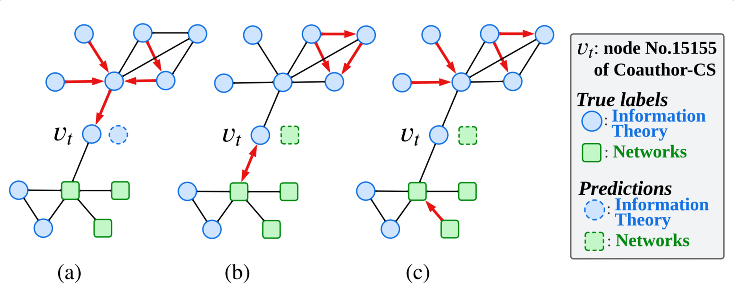
## 4. Experiments

- **Prediction performance (accuracy)**: SUNNY-GNN outperforms all baselines by 3.5% on average.

| | Citeseer | Cora | Pubmed | Amazon | Coauthor-CS | Coauthor-Physics |
|---|---|---|---|---|---|---|
| GCN | 69.84±0.7 | 81.20±0.7 | 77.68±0.7 | 90.18±0.3 | 83.52±0.4 | 92.46±0.2 |
| + GSAT | **70.90±1.1** | 81.48±0.7 | 77.44±0.3 | 88.36±1.3 | 83.76±0.6 | 92.14±0.5 |
| + CAL | 65.60±1.1 | 75.72±1.2 | 73.66±0.8 | 84.32±1.7 | 82.12±1.2 | 91.26±0.7 |
| + SE-GNN | 68.90±0.9 | 80.72±0.1 | 77.56±0.3 | - | 83.14±0.8 | - |
| + ProtGNN | 66.30±2.1 | 77.48±8.7 | 74.18±3.3 | 82.46±1.4 | 79.50±3.7 | 88.80±3.3 |
| + SUNNY-GNN | 70.72±0.8 | **81.68±0.9** | **78.68±0.2** | **90.43±0.4** | **85.03±1.1** | **93.10±0.8** |
| Average impro. (%) | 3.6 ↑ | 3.1 ↑ | 3.4 ↑ | 4.8 ↑ | 3.2 ↑ | 3.0 ↑ |
| GAT | 69.68±1.2 | 81.22±0.7 | 77.50±0.4 | 89.08±1.8 | 84.42±0.8 | 92.30±0.5 |
| + GSAT | 69.42±0.8 | 81.20±0.7 | 77.04±0.3 | 89.73±0.4 | 84.37±0.7 | 91.90±0.8 |
| + CAL | 67.64±1.5 | 76.64±1.1 | 74.74±0.7 | 84.86±11.5 | 78.69±3.8 | 78.24±5.1 |
| + SE-GNN | 68.18±1.1 | 79.46±0.4 | 75.88±0.4 | - | 83.71±0.5 | - |
| + ProtGNN | 69.90±1.5 | 80.40±0.9 | 76.84±0.8 | 86.52±0.3 | 80.95±1.2 | 90.42±2.3 |
| + SUNNY-GNN | **71.30±0.7** | **82.18±1.3** | **78.14±0.3** | **90.78±0.4** | **85.13±0.5** | **93.06±0.6** |
| Average impro. (%) | 3.2 ↑ | 3.1 ↑ | 2.3 ↑ | 3.0 ↑ | 3.4 ↑ | 6.0 ↑ |

- **Explainability performance (fidelity)**: SUNNY-GNN outperforms the baselines by 13.1% on average.

| | Citeseer | |
|---|---|---|
| | $fid_+ \uparrow$ | $fid_- \downarrow$ |
| GCN | | |
| + GNNExplainer | 72.27±4.2 | 9.31±3.4 |
| + PGExplainer | 82.09±7.2 | 0.92±2.6 |
| + ReFine | 83.01±7.1 | 0.78±0.5 |
| + GSAT | 86.75±5.7 | 2.72±1.1 |
| + CAL | 86.44±4.3 | 12.25±3.9 |
| + SUNNY-GNN | **87.29±5.3** | **0.25±0.4** |
| GAT | | |
| + GNNExplainer | 52.95±16.9 | 8.61±10.1 |
| + PGExplainer | 76.80±0.3 | 1.71±2.8 |
| + ReFine | 77.78±2.8 | **0.32±1.3** |
| + GSAT | 72.52±8.1 | 1.55±1.3 |
| + CAL | 77.78±6.5 | 11.46±1.4 |
| + SUNNY-GNN | **79.25±2.4** | 0.46±0.5 |




*Case studies: (a) SUNNY-GNN highlights the most salient input information, while (b) GSAT and (c) CAL fail to.*

## More Information

- **Code**: https://github.com/SJTU-Quant/SUNNY-GNN
- **Contact us**: jialedeng@sjtu.edu.cn