**Algorithm 1:** GNNs with the CIT mechanism

```
Input      : Graph $G = (\mathbf{A}, \mathbf{X})$, label $\mathbf{Y}$
Params     : the probability of transfer $p$, the epochtimes $k$,
             the number of clusters $m$, total iterations $T$
Initialize : GNN model $f_{GNN}$,
             classifier $f_\theta$ (usually the last layer of GNN)
Output     : GNN model $f_{GNN}$, classifier $f_\theta$
for $epoch = 1$ to $T$ do
    Node representation $\mathbf{Z}^{(l)}$ from Eq. (1)
    Cluster representation $\mathbf{H}^c$ from Eq. (2) and Eq. (6)
    Clustering loss $\mathcal{L}_u$ from Eq. (5)
    if $epoch \ \% \ k == 0$ then
        Randomly sample $n \times p$ nodes to calculate Eq. (9)
        Get the new representation $\mathbf{Z}'^{(l)}$
    else
        Keep the node representation $\mathbf{Z}'^{(l)} \leftarrow \mathbf{Z}^{(l)}$
    end
    Classification loss $\mathcal{L}_f$ from Eq. (12)
    Update $f_{GNN}$, $f_\theta$ with Eq. (13)
end
```

# A More details of Section 3

## A.1 Three-Fold optimization

In this section, we detail the process of our CIT mechanism in three-fold optimization, shown in Algorithm1.

## A.2 Computational complexity

Our CIT mechanism has two parts of computation: Clustering process and Cluster Information Transfer process. Let $N$ represent the number of nodes, and $K$ represent the number of clusters. The computational complexity of clustering process is $\mathcal{O}(N^2K + NK^2) = \mathcal{O}(NK(N + K))$. Since the adjacency matrix is usually sparse, the computational complexity can be reduced to $\mathcal{O}(EK)$, where $E$ is the number of non-zero edges in the adjacency matrix. The computational complexity of Cluster Information Transfer process is $\mathcal{O}(pN)$, where $p$ is the probability of transfer. So the computational complexity is $\mathcal{O}(K(E + NK) + pN)$. The space complexity which depends on the dimension of the assignment matrix is $\mathcal{O}(NK)$.

## A.3 Proof of Theorem1

Specifically, we simplify $\mathbf{Z}^{(l)}$ in Eq. (1) as $\mathbf{Z}$, and assume that label $Y \in \{0, 1\}$. There are two clusters $e \in \{D, R\}$. We give the statistics of data. The mean of node representations in cluster $D$ is $E(Z|e = D) = \mu_D$, and the variance of node representations in cluster $D$ is $Var(Z|e = D) = \Sigma_D^2$. Similarly, $E(Z|e = R) = \mu_D$, $Var(Z|e = R) = \Sigma_R^2$. We define the probability of label 0 as $\pi_0 = P(Y = 0)$, label 1 as $\pi_1 = P(Y = 1)$, and then the cluster probability $\pi_D = P(e = D)$ and $\pi_R = P(e = R)$. The conditional probability of label given cluster $D$ as $\pi_{0|D} = P(Y = 0|e = D)$, $\pi_{1|D} = P(Y = 1|e = D)$ and the conditional probability of label given cluster $R$ as $\pi_{0|R} = P(Y = 0|e = R)$, $\pi_{1|R} = P(Y = 1|e = R)$. For analysis, we use $E(Z|Y = 1) = \mu_1$ to represent the mean of node representations with label 1, and $E(Z|Y = 0) = \mu_0$ to represent the mean of node representations with label 0. We assume that there is statistic spurious correlation between clusters and labels, *i.e.*, all label information can be obtained through the label information in each cluster as $\frac{\mu_D}{\pi_{1|D}} + \frac{\mu_R}{\pi_{1|R}} = \mu_1$. At first, we calculate the form of the decision boundary through the data statistics given above and find that the label distribution in clusters affects the decision boundary. And then, we make our transfer on the original data and find that the label distribution in clusters has less influence on it.

*Proof.* Firstly, we use the statistics of node representations in different clusters and clusters probability to calculate the variance:

$$\begin{aligned} Var(Z) &= E(Z^2|e=D)\pi_D + E(Z^2|e=R)\pi_R - E(Z)^2 \\ &= (Var(Z|e=D) + E(Z|e=D)^2)\pi_D \\ &\quad + (Var(Z|e=R) + E(Z|e=R)^2)\pi_R - E(Z)^2 \\ &= (\Sigma_D^2 + \mu_D^2)\pi_D + (\Sigma_R^2 + \mu_R^2)\pi_R - (\mu_D\pi_D + \mu_R\pi_R)^2. \end{aligned} \tag{14}$$

Then we calculate the covariance of $Z$ and $Y$ based on the correlation assumption:

$$\begin{aligned} Cov(Z,Y) &= E(ZY) - E(Y)E(Z) - E(Y)E(Z) + E(Z)E(Y) \\ &= E[E(ZY|Y) - YE(Z) - E(Y)E(Z|Y) + E(X)E(Y)] \\ &= E[(E(Z|Y) - E(E(Z|Y)))(Y - E(Y))] \\ &= Cov((\frac{\mu_D}{\pi_{1|D}} + \frac{\mu_R}{\pi_{1|R}} - \frac{\mu_D}{\pi_{0|D}} - \frac{\mu_R}{\pi_{0|R}})Y, Y) \\ &= (\frac{\mu_D}{\pi_{1|D}} + \frac{\mu_R}{\pi_{1|R}} - \frac{\mu_D}{\pi_{0|D}} - \frac{\mu_R}{\pi_{0|R}})\pi_0\pi_1. \end{aligned} \tag{15}$$

Combining Eq. (14) and Eq. (15) we can see that, the label distribution in cluster $\pi_{Y|e}$ affects the covariance $Cov(Z,Y)$. So in this case, the decision boundary is directly influenced by cluster information. □

### A.4 Proof of Theorem2

*Proof.* For analysis, we assume that there are $n_D$ nodes belonging to cluster $D$ and $n_R$ nodes belonging to cluster $R$. So after the transfer, the probability of cluster $D$ is $\pi'_D = \frac{n_D + n_R p}{n_D + n_R}$ and probability of cluster $R$ is $\pi'_R = \frac{n_R - n_R p}{n_D + n_R}$. The new variance can be calculated as follows:

$$\begin{aligned} Var(Z) &= (\Sigma_D'^2 + \mu_D^2)\pi'_D + (\Sigma_R'^2 + \mu_R^2)\pi'_R - (\mu_D\pi'_D + \mu_R\pi'_R)^2 \\ &= (\frac{\Sigma_D^2 n_D}{n_D + n_R p} + \frac{n_D}{n_D + n_R p}\mu_D^2 + \frac{n_R p}{n_D + n_R p}\mu_R^2 \\ &\quad - (\frac{n_D}{n_D + n_R p}\mu_D + \frac{n_R p}{n_D + n_R p}\mu_R)^2 + \mu_D^2)(\frac{n_D + n_R p}{n_D + n_R}) \\ &\quad + (\Sigma_R^2 + \mu_R^2)\frac{n_R - n_R p}{n_D + n_R} - (\mu_D\frac{n_D + n_R p}{n_D + n_R} + \mu_R\frac{n_R - n_R p}{n_D + n_R})^2. \end{aligned} \tag{16}$$

We use $\pi'_{Y|e}$ to represent new label distribution in each clusters. Then the new covariance can be represented as follows:

$$Cov(Z,Y) = (\frac{\mu_D}{\pi'_{1|D}} + \frac{\mu_R}{\pi'_{1|R}} - \frac{\mu_D}{\pi'_{0|D}} - \frac{\mu_R}{\pi'_{0|R}})\pi_0\pi_1. \tag{17}$$

From Eq. (16) and Eq. (17) we can see, the label distribution in cluster still have no effect on $Var(Z)$. So we analyze the $\pi'_{Y|e}$ which affects the $Cov(Z,Y)$. We take cluster $D$ as an example. We use $n_{D0}$ and $n_{R0}$ to represent the number of nodes with label 0 in each cluster. Similarly, $n_{D1}$ and $n_{R1}$ to represent the number of nodes with label 1 in each cluster. After the transfer, we calculate the probability of new label-0 in cluster $\pi'_{0|D} = \frac{n_{D0} + p n_R \pi_{0|R}}{n_D + n_R p} = \frac{n_{D0} + p n_{R0}}{n_D + n_R p}$. We can see that the conditional probability $\pi'_{0|D}$ approaches to $\pi_0$, which is as same as label 1, meaning that the effect

Table 3: Data statistics.

| Datasets | Nodes | Edges | Features | Classes | Structures |
|---|---|---|---|---|---|
| Cora | 2708 | 5429 | 1433 | 7 | 1 |
| Citeseer | 3327 | 4732 | 3703 | 6 | 1 |
| Pubmed | 19717 | 44324 | 500 | 3 | 1 |
| ACM | 3025 | 29281 | 1830 | 3 | PAP |
| | | 2210761 | | | PSP |
| IMDB | 3550 | 66428 | 1007 | 3 | MAM |
| | | 13788 | | | MDM |
| Twitch-Explicit | 9498 | 153138 | 3170 | 2 | DE |
| | 7126 | 35324 | | | ENGB |
| | 4648 | 59382 | | | ES |
| | 6549 | 1123666 | | | FR |
| | 4385 | 37304 | | | RU |
| | 2772 | 63462 | | | TW |

between the decision boundary of classifier and cluster information is weakened. When $p = 1$, $\pi'_{0|D} = \pi_0$ and $\pi'_{1|D} = \pi_1$. In this case, the $Cov(Z, Y) = \mu_D(\pi_1 - \pi_0)$, which has no relations about cluster information. □

## B More details of Section 4

### B.1 Data statistics

- **Cora** [20]: The Cora is a citation network. The nodes represent papers and are classified into three classes. The edges represent their citation relationships. Node attributes are bag-of-words representations of the papers and the nodes are labeled based on the paper topics.
- **Citeseer** [20]: The Citeseer is a link dataset bulit from citeseer web dataset. The nodes are publications and are divided into six areas. Node attributes are representations of the papers. The edges are citation links.
- **Pubmed** [30]: The Pubmed is a searchable database in the medical field. It consists of nearly twenty thousand nodes. All nodes are divided into three classes. Edges represent papers citation relationship. Node attributes are bag-of-words of the papers.
- **ACM** [27]: This network is extracted from ACM dataset where nodes represent papers and there is an edge between two papers if they have the same author or same subject. So the nodes have two relations which are Papers-Authors-Papers (PAP) and Papers-Subject-Papers (PSP). All the papers are divided into three classes. The features are the bag-of-words representations of paper keywords.
- **IMDB** [27]: IMDB is a movie network dataset where nodes represent movies and there is an edge between two movies if they have the same director or same actor. So the nodes have two relations which are Movie-Actor-Movie (MAM) and Movie-Director-Movie (MDM). All the movies are divided into three classes and features are the bag-of-words of reviews and movie information.
- **Twitch-Explicit** [17]: Twitch datasets contain several networks where nodes represent Twitch users and edges represent their mutual friendships. Each network is collected from a particular region. Different networks have different size, densities and maximum node degrees. All nodes are divided into two classes.

### B.2 Additional results

For more comparison, we show result of deleting edges in Table 4.

Table 4: Quantitative results (%±$\sigma$) on node classification for perturbation on graph structures data while the superscript refers to the results of paired t-test (* for 0.05 level and ** for 0.01 level).

| Method | Dele-0.2 | | | | | |
|---|---|---|---|---|---|---|
| | Cora | | Citeseer | | Pubmed | |
| | Acc | Macro-f1 | Acc | Macro-f1 | Acc | Macro-f1 |
| GCN | 80.04±0.48 | 78.86±0.62 | 69.68±0.38 | 67.29±0.49 | 77.48±0.71 | 77.32±0.65 |
| SR-GCN | 79.80±0.61 | 78.31±0.55 | 70.03±0.87 | 67.62±0.80 | 78.10±1.10 | 77.63±1.21 |
| EERM-GCN | 78.57±0.78 | 76.32±0.81 | 69.95±0.42 | 67.97±0.60 | - - | |
| CIT-GCN(w/o) | 80.36±0.34 | 79.03±0.40 | **71.38±0.38**$^{**}$ | **68.61±0.38**$^{**}$ | **79.38±0.37**$^{**}$ | **78.85±0.34**$^{**}$ |
| CIT-GCN | **80.70±0.42** | **79.67±0.51** | 71.20±0.51 | 68.52±0.46 | 78.40±0.62 | 77.96±0.50 |
| GAT | 80.22±0.41 | 79.27±0.35 | 69.10±0.46 | 66.20±0.39 | 76.35±0.62 | **75.95±0.58** |
| SR-GAT | 80.25±0.65 | 79.29±0.57 | 68.80±0.49 | 66.28±0.32 | **76.55±0.47** | 75.39±0.56 |
| EERM-GAT | 79.15±0.38 | 77.92±0.29 | 68.15±0.37 | 65.31±0.45 | - - | |
| CIT-GAT(w/o) | 80.98±0.60 | 80.07±0.46 | 69.98±0.62 | 67.32±0.67 | 76.21±0.48 | 75.12±0.56 |
| CIT-GAT | **81.35±0.35**$^{*}$ | **80.27±0.44**$^{*}$ | **70.11±0.47**$^{*}$ | **67.77±0.57**$^{*}$ | 76.05±0.54 | 75.65±0.46 |
| APPNP | 80.84±0.54 | 80.13±0.61 | 70.62±0.96 | 67.86±0.64 | 79.41±0.37 | 78.87±0.36 |
| SR-APPNP | 80.11±0.65 | 80.06±0.77 | 69.27±0.43 | 67.77±0.39 | 75.85±0.55 | 75.43±0.58 |
| EERM-APPNP | 79.17±0.77 | 79.72±0.59 | 71.30±0.61 | 67.92±0.57 | - - | |
| CIT-APPNP(w/o) | **81.46±0.40** | 80.57±0.47 | **72.06±0.28**$^{**}$ | **69.01±0.32**$^{**}$ | 79.35±0.52 | 78.68±0.51 |
| CIT-APPNP | 81.43±0.39 | **80.78±0.44** | 71.84±0.51 | 68.57±0.55 | **79.88±0.37** | **79.29±0.46** |
| GCNII | 82.82±0.48 | 81.03±0.47 | 71.58±0.50 | 68.24±0.61 | 78.65±0.64 | 77.92±0.53 |
| SR-GCNII | 81.75±0.41 | 81.09±0.38 | 70.24±0.76 | 66.87±0.83 | 78.10±0.52 | 76.76±0.61 |
| EERM-GCNII | 80.05±0.67 | 79.12±0.53 | 71.11±0.63 | 68.02±0.79 | - - | |
| CIT-GCNII(w/o) | 82.41±0.43 | 81.07±0.35 | 71.70±0.92 | 68.56±0.88 | 78.85±0.33 | **79.19±0.23**$^{*}$ |
| CIT-GCNII | **83.20±0.58** | **81.70±0.63** | **72.38±0.62**$^{*}$ | **69.13±0.31**$^{*}$ | **79.80±0.73**$^{*}$ | 79.17±0.66 |
| | Dele-0.5 | | | | | |
| GCN | 77.28±0.47 | 75.30±0.56 | 68.52±0.33 | 65.59±0.36 | 77.04±0.32 | 76.64±0.38 |
| SR-GCN | 76.70±0.81 | 74.59±0.67 | 67.72±1.10 | 64.58±1.22 | 76.35±0.56 | 76.54±0.63 |
| EERM-GCN | 77.30±0.31 | 75.18±0.45 | 68.65±0.45 | 65.55±0.36 | - - | |
| CIT-GCN(w/o) | 77.05±0.47 | 75.17±0.38 | 70.02±0.49 | 67.10±0.44 | 77.83±0.21 | **77.63±0.36**$^{*}$ |
| CIT-GCN | **77.50±0.51** | **75.58±0.66** | **70.12±0.55**$^{**}$ | **66.81±0.56**$^{**}$ | **77.90±0.46** | 77.23±0.53 |
| GAT | 77.22±0.37 | 75.81±0.32 | 68.94±0.47 | 65.98±0.55 | 75.92±0.63 | 75.61±0.66 |
| SR-GAT | 77.38±0.42 | 75.86±0.43 | 68.27±0.73 | 64.24±0.92 | 75.31±0.67 | 74.24±0.78 |
| EERM-GAT | 76.62±0.73 | 74.38±0.68 | 67.12±0.54 | 64.01±0.62 | - - | |
| CIT-GAT(w/o) | 77.52±0.46 | 76.07±0.42 | 69.38±0.57 | 66.24±0.59 | 76.01±0.47 | 75.97±0.46 |
| CIT-GAT | **77.72±0.55** | **76.43±0.57** | **69.44±0.56**$^{*}$ | **66.58±0.47**$^{*}$ | **76.79±0.77**$^{*}$ | **76.43±0.67**$^{*}$ |
| APPNP | 78.52±0.66 | 77.12±0.67 | 69.41±0.63 | 66.43±0.60 | 77.80±0.63 | 77.43±0.61 |
| SR-APPNP | 77.55±0.49 | 76.97±0.42 | 70.81±0.47 | 66.78±0.61 | 76.45±0.51 | 76.37±0.58 |
| EERM-APPNP | 77.31±0.55 | 76.87±0.61 | 69.91±0.59 | 66.32±0.62 | - - | |
| CIT-APPNP(w/o) | 78.80±0.53 | 77.31±0.43 | 70.41±0.42 | 67.30±0.39 | **78.08±0.34** | **77.73±0.32** |
| CIT-APPNP | **79.02±0.52** | **78.27±0.48** | **71.06±0.55**$^{*}$ | **67.57±0.58**$^{*}$ | 77.60±0.61 | 77.38±0.53 |
| GCNII | 80.48±0.45 | 78.65±0.39 | 70.04±0.89 | 66.61±0.83 | 78.40±0.68 | 78.18±0.78 |
| SR-GCNII | 80.03±0.60 | 78.38±0.53 | 70.19±0.71 | 67.01±0.82 | 77.98±1.01 | 76.89±0.92 |
| EERM-GCNII | 78.52±0.82 | 77.02±0.93 | 69.40±0.67 | 66.81±0.91 | - - | |
| CIT-GCNII(w/o) | 79.84±0.43 | 78.04±0.47 | 70.72±0.69 | 67.41±0.57 | **78.58±0.38** | **78.24±0.44** |
| CIT-GCNII | **80.58±0.62** | **78.94±0.45** | **71.32±0.44**$^{*}$ | **68.04±0.33**$^{*}$ | 78.13±0.61 | 77.89±0.70 |

### B.3 Implementation details

For every GNNs method, we follow the parameter settings from their original paper. SR-GNN and EERM-GNN are initialized with same parameters suggested by their papers and we also further carefully turn parameters to get optimal performance.

For our CIT-GNN, we do not change the parameters of the previous part of GNN backbones and only make an adjustment on our module. Although our transfer process is conducted every $k$ epochs, the clustering process proceeds all the training procedure. For GCN, GAT and GCNII, we put our CIT mechanism before the last layer of GNN. For APPNP, we put it in features extract process, that is, before the last layer of linear transform. We search on the probability of transfer $p$ from 0.05 to 0.3 with step 0.05 and tune epochtimes $k$ of CIT from 2 to 50. For dropout rate, we test ranging is from 0.1 to 0.6. Moreover, we tune the numbers of clusters which is the parameter from spectral clustering from [10, 20, 30, 40, 50, 100, 200]. We set classification loss coefficient, cutloss coefficient and orthogonality loss coefficient as 0.5, 0.3, 0.2 respectively. For all models, we randomly run 5 times and report the average results. For every dataset, we only use original attributes of target nodes, and assign one-hot id vectors to nodes of other types. We report our experiment setting and parameters in supplement.

#### B.3.1 Experiment settings

All experiments are conducted with the following setting:

- Operating system: CentOS Linux release 7.6.1810
- CPU: Intel(R) Xeon(R) CPU E5-2620 v4 @ 2.10GHz
- GPU: GeForce RTX 2080 Ti with 11GB and GeForce RTX 3090 with 24GB
- Software versions: Python 3.8; Pytorch 1.10.1; Cuda 11.1;

### B.3.2 Baselines

The publicly available implementations of Baselines can be found at the following URLs:

- GCN: `https://github.com/tkipf/pygcn`
- GAT: `https://github.com/Diego999/pyGAT`
- APPNP: `https://github.com/gasteigerjo/ppnp`
- GCNII: `https://github.com/chennnM/GCNII`
- SR-GNN: `https://github.com/GentleZhu/Shift-Robust-GNNs`
- EERM: `https://github.com/qitianwu/GraphOOD-EERM`

For a fairly comparison, we plug the three methods in same code of GNNs model referred from their papers.

### B.3.3 Hyper parameter settings

Our CIT-GNN contains four hyper-parameter, the probability of transfer $p$, epochtimes $k$, the number of clusters $m$ and $dropout$.

### B.3.4 Settings for Section Perturbation on graph structures data

For Cora, Citeseer and Pubmed, our hyper-parameter settings are as follows respectively:

- CIT-GCN: $p$=0.2/0.1/0.02, $k$=5/5/20,
  $m$=100/20/100, $dropout$=0.5/0.1/0.5 .
- CIT-GAT: $p$=0.1/0.1/0.02, $k$=5/5/5,
  $m$=100/20/100, $dropout$=0.6/0.5/0.3 .
- CIT-APPNP: $p$=0.2/0.2/0.02, $k$=20/20/20,
  $m$=200/10/200, $dropout$=0.6/0.1/0.5 .
- CIT-GCNII: $p$=0.1/0.02/0.1, $k$=5/10/20,
  $m$=100/40/200, $dropout$=0.5/0.3/0.3 .

### B.3.5 Settings for Section Multiplex networks data

For ACM and IMDB (two relations), our hyper-parameter settings are as follows respectively:

- CIT-GCN: $p$=0.2/0.02/0.1/0.05, $k$=10/5/20/5,
  $m$=10/100/40/200, $dropout$=0.5/0.3/0.6/0.3 .
- CIT-GAT: $p$=0.2/0.1/0.2/0.1, $k$=10/5/5/5,
  $m$=10/50/40/50, $dropout$=0.1/0.1/0.1/0.5 .
- CIT-APPNP: $p$=0.2/0.1/0.1/0.1, $k$=5/5/5/5,
  $m$=10/20/40/40, $dropout$=0.5/0.5/0.5/0.3 .
- CIT-GCNII: $p$=0.02/0.1/0.1/0.1, $k$=5/5/5/20,
  $m$=40/100/100/200, $dropout$=0.1/0.3/0.3/0.1 .

### B.3.6 Settings for Section Multigraph data

- CIT-GCN: $p$=0.02, $k$=20, $m$=200, $dropout$=0.3 .
- CIT-GAT: $p$=0.05, $k$=20, $m$=200, $dropout$=0.3 .
- CIT-APPNP: $p$=0.02, $k$=5, $m$=200, $dropout$=0.3 .
- CIT-GCNII: $p$=0.05, $k$=20, $m$=200, $dropout$=0.5 .