A Appendix

This appendix provides the proofs of the theorems, and a comparison of running time.

A.1 Proof of Theorem 1

Since there are two groups of variables in Eq. (9), i.e., the node representations $\mathbf{H}$ and the learned topology $\mathbf{C}$, Eq. (9) can be minimized with respective to one group of variables by fixing the other group. When the graph topology $\mathbf{C}$ is fixed, minimizing Eq. (9) is equivalent to minimizing the following objective w.r.t. the node representations $\mathbf{H}$, as

$$\min_{\mathbf{H}} tr(\mathbf{H}^T \mathbf{L}_C \mathbf{H}). \tag{17}$$

According to [26, 27, 28], the gradient descent method employed to minimize Eq. (17) is identical to the graph convolutional operation in GCN, i.e., Eq. (2) with a fixed c_{uv}, by ignoring the mapping function $\mathbf{W}$ and the nonlinear function $\sigma(\cdot)$.

When the node representations $\mathbf{H}$ are fixed, minimizing Eq. (9) is equivalent to minimizing the following objective w.r.t. node representations $\mathbf{C}$, as

$$\min_{\mathbf{C},\mathbf{H}} \sum_{u,v} \left(b_{uv} c_{uv} + \gamma c_{uv}^2\right) + 2tr(\mathbf{H}^T \mathbf{L}_C \mathbf{H}), \tag{18}$$

$$s.t.\ \forall u \sum c_{uv} = 1,\ 0 \le c_{uv} \le 1. \tag{19}$$

Eq. (18) can then be rewritten as

$$\min_{\mathbf{C}} \sum_{u,v} \left(b_{uv} c_{uv} + \gamma c_{uv}^2\right) + \sum_{u,v} c_{uv} ||\mathbf{h}_u - \mathbf{h}_v||^2. \tag{20}$$

If we let

$$o_{uv} = b_{uv} + ||\mathbf{h}_u - \mathbf{h}_v||, \tag{21}$$

Eq. (18) is equivalent to

$$\min_{\mathbf{c}_u^T \mathbf{1}=1,\ 0 \le \mathbf{c}_u \le 1} \left\| \mathbf{c}_u + \frac{1}{2\gamma} \mathbf{o}_u \right\|^2, \tag{22}$$

where $\mathbf{c}_u$ and $\mathbf{o}_u$ are the vectors containing c_{uv} and o_{uv}, respectively. Eq. (22) can be minimized by applying the Lagrangian multiplier method and KKT condition. Then, the solution to Eq. (18) is

$$c_{uv} = \left(-\frac{o_{uv}}{2\gamma} + \eta\right)_+ = ReLU\left(-\frac{o_{uv}}{2\gamma} + \eta\right), \tag{23}$$

where η is one of the Lagrangian multipliers. The tuning of η makes $\mathbf{c}_u^T \mathbf{1} = 1$.

Let b_{uv} in Eq. (9) be the combination of topology and node attributes, as

$$b_{uv} = -\zeta a_{uv} \left(\mathbf{w}^T [\mathbf{x}_u || \mathbf{x}_v]\right), \tag{24}$$

where $[\mathbf{x}_u || \mathbf{x}_v]$ denotes the concatenation of $\mathbf{x}_u$ and $\mathbf{x}_v$, and $\mathbf{w}$ is the learnable parameters, which with the same length as $[\mathbf{x}_u || \mathbf{x}_v]$. a_{uv} stands for the corresponding element in the adjacency matrix. ζ denotes the importance of this term. Then, Eq. (23) can be revised to

$$c_{uv} = ReLU\left(\frac{\zeta}{2\gamma} a_{uv} \left(\mathbf{w}^T [\mathbf{x}_u || \mathbf{x}_v]\right) - \frac{||\mathbf{h}_u - \mathbf{h}_v||}{2\gamma}\right). \tag{25}$$

If the importance parameter ζ is large, Eq. (25) can be simplified as

$$c_{uv} = ReLU\left(\frac{\zeta}{2\gamma} a_{uv} \left(\mathbf{w}^T [\mathbf{x}_u || \mathbf{x}_v]\right)\right). \tag{26}$$

It can be observed that $c_{uv} \neq 0$ only if $a_{uv} \neq 0$, i.e., nodes v and u are connected. Thus, the learned topology in Eq. (26) is similar to that in GAT [18].

Therefore, the Uniform Message Passing in Eq. (2) with learnable weights is essentially the gradient descent of the objective function in Eq. (9).

A.2 Proof of Theorem 4

Note that the rank of a matrix $\mathbf{L}_C \in \mathbf{R}^{N\times N}$ is the difference between N and its multiplier with an eigenvalue 0, i.e., $rank(\mathbf{L}_C) = N - F$. As shown in Theorem 3, the multiplier F with the eigenvalue 0 for the Laplacian matrix $\mathbf{L}_C$ equals to the number of connected components in the graph with the similarity matrix $\mathbf{C}$. Then, the constraint $rank(\mathbf{L}_C) = N - F$ is equivalent to partitioning the graph into F connected components. Besides, to constrain the multiplier with the eigenvalue 0 being F is equivalent to minimizing F smallest eigenvalues, i.e, $\sum_{f=1}^{F}\sigma_f(\mathbf{L}_C)$. According to Theorem 2, $\sum_{f=1}^{F}\sigma_f(\mathbf{L}_C)$ is the minima of $tr(\mathbf{H}^T\mathbf{L}_C\mathbf{H})$. Therefore, Eq. (9) in Theorem 1 is equivalent to Eq. (12) in Theorem 4, and the Uniform Message Passing in Eq. (2) actually partitions graph into F connected components by learning the topology $\mathbf{C}$.

A.3 Proof of Theorem 5

Most of this proof can be deducted in a similar approach to the proof of Theorem 4. The remaining part is only presented to prove that each learned graph actually partitions the graph into two connected components, i.e., the constraint $rank(\mathbf{L}_C^{(f)}) = N - 2$.

According to Theorem 3, to partition graph into two connected components, the multiplier with the eigenvalue 0 should be 2, and $\sigma_1(\mathbf{L}_C) + \sigma_2(\mathbf{L}_C)$ should be accordingly minimized. In fact, $\sigma_1(\mathbf{L}_C) = 0$ for all the Laplacian matrix $\mathbf{L}_C$. Thus, to partition the graph into two connected components, $\sigma_2(\mathbf{L}_C)$ should be minimized. According to [32], it holds that

$$\sigma_2(\mathbf{L}_C) = \inf_{\mathbf{h}\perp\mathbf{1}} \frac{\mathbf{h}^T\mathbf{L}_C\mathbf{h}}{\mathbf{h}^T\mathbf{h}}. \tag{27}$$

In the following paragraphs, we prove that the following approximation

$$\inf_{\mathbf{h}\neq\mathbf{1},||\mathbf{h}||_2=1} \mathbf{h}^T\mathbf{L}_C\mathbf{h}, \tag{28}$$

is bounded by $\sigma_2(\mathbf{L}_C)$.

For each $\mathbf{h} \neq \mathbf{1}, ||\mathbf{h}||_2 = 1$ can be decomposed into two mutually perpendicular vectors, i.e.,

$$\mathbf{h} = \alpha\mathbf{1} + \mathbf{g}, \tag{29}$$

where $\mathbf{g}^T\mathbf{1} = 0$. Eq. (29) can be reformed to $\alpha = \frac{\mathbf{h}^T\mathbf{1}}{N}$, which is the average of all the elements in $\mathbf{h}$. Besides, the norm of $\mathbf{g}$ is bounded, i.e., $t_1 \leq ||\mathbf{g}||^2 \leq t_2$. Then, we obtain

$$\begin{aligned}
\mathbf{h}^T\mathbf{L}_C\mathbf{h} &= (\alpha\mathbf{1}+\mathbf{g})^T\mathbf{L}_C(\alpha\mathbf{1}+\mathbf{g}) &(30)\\
&= \alpha^2\mathbf{1}^T\mathbf{L}_C\mathbf{1} + \mathbf{g}^T\mathbf{L}_C\mathbf{g} + 2\alpha\mathbf{1}^T\mathbf{L}_C\mathbf{g} &(31)\\
&= \mathbf{g}^T\mathbf{L}_C\mathbf{g}, &(32)
\end{aligned}$$

and

$$t_1\sigma_2(\mathbf{L}_C) = t_1 \inf_{\mathbf{g}\perp\mathbf{1}} \frac{\mathbf{g}^T\mathbf{L}_C\mathbf{h}}{\mathbf{g}^T\mathbf{g}} \leq \inf \mathbf{h}^T\mathbf{L}_C\mathbf{h} \leq t_2 \inf_{\mathbf{g}\perp\mathbf{1}} \frac{\mathbf{g}^T\mathbf{L}_C\mathbf{h}}{\mathbf{g}^T\mathbf{g}} = t_2\sigma_2(\mathbf{L}_C). \tag{33}$$

Thus, $\inf_{\mathbf{h}\neq\mathbf{1},||\mathbf{h}||_2=1} \mathbf{h}^T\mathbf{L}_C\mathbf{h}$ is bounded by $\sigma_2(\mathbf{L}_C)$. Therefore, if the $\mathbf{H} = \mathbf{h} \in \mathbf{R}^N$ in Eq. (9) is a vector and all elements are not equal, then minimizing $tr(\mathbf{H}^T\mathbf{L}_C\mathbf{H})$ is actually minimizing $\sigma_2(\mathbf{L}_C)$ according to Eq. (33). Thus, it is equivalent to the constraint $rank(\mathbf{L}_C) = N - 2$.

Since the diverse message passing in Eq. (4) is equivalent to learning different graph $\mathbf{C}^{(f)}$ for different attributes f, thus each learned graph $\mathbf{C}^{(f)}$ is to partition graph into two components, i.e., the constraint $rank(\mathbf{L}_C^{(f)}) = N - 2$.

A.4 Running Time

Here we compare the speed of the proposed DMP with the baseline methods designed for networks with heterophily, especially the SOTA method H2GCN. The results shown in Table 4 are the runtimes of training and testing the model in terms of seconds. It can be observed that the proposed DMP is faster than the others on large networks. Note that DMP is obviously faster than H2GCN on all the datasets.

Table 4: Comparison of running time of the methods for networks with heterophily (seconds).

Dataset	Cornell	Texas	Actor	Chameleon	Citeseer	Cora	Pubmed	Squirrel
DMP	18.87	18.53	19.13	20.60	20.70	21.79	18.24	30.39
JKNet	8.44	16.43	9.45	25.16	20.74	22.29	20.99	27.76
ChebNet	10.56	11.41	16.34	31.90	28.95	27.72	42.66	60.54
H2GCN	128.45	83.06	363.03	226.78	116.09	288.23	340.25	869.14

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Experimental Results Analysis
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] See Appendix
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] In the supplemental material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Experimental Setup and Code.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Experimental Setup.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes] Citing the paper containing the data.
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]