APPENDIX

Mengzhe Jia, Hanxi Zhu, Zhengpin Li, Ruyu Wang, Jian Wang

School of Data Science, Fudan University

1. THEORY

In this section, we aim to theoretically analyze how the feature embeddings generated by GRASP—after the graph-splitting and frequency-aware aggregation modules—facilitate correct node classification. For the analysis, we adopt the contextual stochastic block model (cSBM) [1], a generative model for graphs with node features and community structure that is widely used in graph learning studies [2, 3]. To facilitate the theoretical analysis, we adopt the following assumptions. Although they introduce certain simplifications, they are widely used in the literature and do not compromise the general applicability of our results.

Assumption 1. *We limit the classification setting to two classes, i.e., $C = 2$.*

This simplification is common in theoretical analysis [3], as it allows us to focus on the core mechanisms of the model without getting entangled in the additional complexity introduced by multi-class problem. Importantly, many key insights and derivations in binary settings can be naturally extended to multi-class scenarios through techniques like one-vs-rest. Therefore, this assumption does not limit the relevance of the analysis to real-world applications involving more than two classes.

Assumption 2. *The class distribution is balanced, i.e., $\mathbb{P}(\mathbf{Y} = 1) = \mathbb{P}(\mathbf{Y} = 0)$.*

This assumption helps remove potential confounding factors caused by class imbalance, which can skew statistical estimations and model behavior. By ensuring that both classes are equally represented, we can isolate the effect of structural and frequency-aware design choices on model performance. In practice, even if the data is imbalanced, techniques such as re-weighting or re-sampling can adjust the training process to approximate a balanced setting, further supporting the relevance of this assumption.

Assumption 3. *The aggregated feature embeddings $\mathbf{h}_i$, $i \in \{1, \cdots, N\}$ produced by GRASP share the same variance, i.e., each follows a distribution with covariance $\sigma\mathbf{I}$.*

Corresponding author: jian_wang@fudan.edu.cn

This is a standard assumption in theoretical analysis and is motivated by the fact that GRASP applies the same filtering architecture across all nodes, which tends to normalize the variance of the resulting embeddings. Moreover, assuming the same covariance simplifies the mathematical derivation and allows us to focus on the relative positions of embeddings in the latent space. In real-world settings, normalization layers (e.g., batch/layer norm) and regularization techniques often implicitly enforce similar statistical properties, lending further justification to this assumption.

Then, we establish Theorem 1.1 to give the probability to classify nodes with same label to the same class.

Theorem 1.1. *Under the above assumptions, after propagation in GRASP, for any two nodes u and v from the same class c_1, we have:*

$$|\mathbb{P}(\hat{y}_u = c_1|\mathbf{h}_u) - \mathbb{P}(\hat{y}_v = c_1|\mathbf{h}_v)| \leq \frac{1}{\sigma^2} \|\mathbf{h}_u - \mathbf{h}_v\| \cdot \rho. \tag{1}$$

Here, $\mathbb{P}(\hat{y} = c|\mathbf{h})$ denotes the conditional prediction probability of assigning a node to class c given its feature $\mathbf{h}$ given by GRASP . $\rho = \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|$ denotes the feature separation distance between classes, which reflects the discriminability of mean vectors across communities.

Proof. GRASP computes feature representations by aggregating information over homophilic and heterophilic subgraphs, formulated respectively as $\mathbf{H}_{\text{hom}} = \frac{1}{2}(\mathbf{I} + \mathbf{A}_{\text{hom}})\mathbf{X}$ and $\mathbf{H}_{\text{het}} = \frac{1}{2}(\mathbf{I} - \mathbf{A}_{\text{het}})\mathbf{X}$. Then following Gaussian distribtiuon, for any node i in class c_1, we obtain:

$$\mathbf{h}_{i,\text{ hom}} \sim \mathcal{N}(\boldsymbol{\mu}_1, \sigma\mathbf{I}),\ \mathbf{h}_{i,\text{ het}} \sim \mathcal{N}\left(\frac{\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2}{2}, \sigma\mathbf{I}\right). \tag{2}$$

Likewise, for a node j from class c_2:

$$\mathbf{h}_{j,\text{ hom}} \sim \mathcal{N}(\boldsymbol{\mu}_2, \sigma\mathbf{I}),\ \mathbf{h}_{j,\text{ het}} \sim \mathcal{N}\left(\frac{\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1}{2}, \sigma\mathbf{I}\right). \tag{3}$$

GRASP then combines the two representations into the final embedding via a weighted sum $\mathbf{H} = \tau\mathbf{H}_{\text{hom}} + (1 - \tau)\mathbf{H}_{\text{het}}$:

$$\begin{aligned}
\mathbf{h}_i &\sim \mathcal{N}\left(\frac{1+\tau}{2}\boldsymbol{\mu}_1 - \frac{1-\tau}{2}\boldsymbol{\mu}_2,\ \sigma'\mathbf{I}\right), \text{ for } i \text{ belongs to class } c_1,\\
\mathbf{h}_j &\sim \mathcal{N}\left(\frac{1+\tau}{2}\boldsymbol{\mu}_2 - \frac{1-\tau}{2}\boldsymbol{\mu}_1,\ \sigma'\mathbf{I}\right), \text{ for } j \text{ belongs to class } c_2,
\end{aligned} \tag{4}$$

where $\sigma' = \sqrt{\tau^2 + (1-\tau)^2}\sigma$. Here we denote $\boldsymbol{\mu}_1' = \frac{1+\tau}{2}\boldsymbol{\mu}_1 - \frac{1-\tau}{2}\boldsymbol{\mu}_2$, $\boldsymbol{\mu}_2' = \frac{1+\tau}{2}\boldsymbol{\mu}_2 - \frac{1-\tau}{2}\boldsymbol{\mu}_1$.

The conditional probability that node u from class c_1 is predicted as c_1 by GRASP, given its embedding $\mathbf{f}_u$, can be computed via Bayes' rule:

$$\mathbb{P}(\hat{y} = c_1|\mathbf{h}_u) = \frac{\mathbb{P}(\mathbf{h}_u|\hat{y}_u = c_1)\mathbb{P}(\hat{y}_u = c_1)}{\mathbb{P}(\mathbf{h}_u|\hat{y}_u = c_1)\mathbb{P}(\hat{y}_u = c_1) + \mathbb{P}(\mathbf{h}_u|\hat{y}_u = c_2)\mathbb{P}(\hat{y}_u = c_2)}. \tag{5}$$

Since we assume a balanced class prior with $\mathbb{P}(\mathbf{Y} = 1) = \mathbb{P}(\mathbf{Y} = 0)$ as Assumption 2 shows,

$$\begin{aligned}\mathbb{P}(\hat{y} = c_1|\mathbf{h}_u) &= \frac{\mathbb{P}(\mathbf{h}_u|\hat{y}_u = c_1)}{\mathbb{P}(\mathbf{h}_u|\hat{y}_u = c_1) + \mathbb{P}(\mathbf{h}_u|\hat{y}_u = c_2)} \\ &= \frac{\exp\left(-\frac{(\mathbf{h}_u - \boldsymbol{\mu}_1')^2}{2\sigma'^2}\right)}{\exp\left(-\frac{(\mathbf{h}_u - \boldsymbol{\mu}_1')^2}{2\sigma'^2}\right) + \exp\left(-\frac{(\mathbf{h}_u - \boldsymbol{\mu}_2')^2}{2\sigma'^2}\right)}.\end{aligned} \tag{6}$$

To make the equation more compact and readable, we define

$$\begin{aligned}\mathbf{z}_{u1} &= \left(\mathbf{h}_u - \boldsymbol{\mu}_1'\right)^2, \ \mathbf{z}_{u2} = \left(\mathbf{h}_u - \boldsymbol{\mu}_2'\right)^2, \\ \mathbf{z}_{v1} &= \left(\mathbf{h}_v - \boldsymbol{\mu}_1'\right)^2, \ \mathbf{z}_{v2} = \left(\mathbf{h}_v - \boldsymbol{\mu}_2'\right)^2,\end{aligned} \tag{7}$$

and rewrite the probability expression (Eq. 6) accordingly as

$$\mathbb{P}(\hat{y} = c_1|\mathbf{h}_u) = \frac{\exp\left(-\frac{\mathbf{z}_{u1}}{2\sigma'^2}\right)}{\exp\left(-\frac{\mathbf{z}_{u1}}{2\sigma'^2}\right) + \exp\left(-\frac{\mathbf{z}_{u2}}{2\sigma'^2}\right)}. \tag{8}$$

Consequently, the difference in classification probability between nodes u and v from the same class is:

$$\begin{aligned}&|\mathbb{P}(\hat{y}_u = c_1|\mathbf{h}_u) - \mathbb{P}(\hat{y}_v = c_1|\mathbf{h}_v)| \\ &= \left|\frac{\exp\left(-\frac{\mathbf{z}_{u1}}{2\sigma'^2}\right)}{\exp\left(-\frac{\mathbf{z}_{u1}}{2\sigma'^2}\right) + \exp\left(-\frac{\mathbf{z}_{u2}}{2\sigma'^2}\right)} - \frac{\exp\left(-\frac{\mathbf{z}_{v1}}{2\sigma'^2}\right)}{\exp\left(-\frac{\mathbf{z}_{v1}}{2\sigma'^2}\right) + \exp\left(-\frac{\mathbf{z}_{v2}}{2\sigma'^2}\right)}\right| \\ &= \left|\frac{\exp\left(-\frac{\mathbf{z}_{u1}}{2\sigma'^2}\right)\exp\left(-\frac{\mathbf{z}_{v2}}{2\sigma'^2}\right) - \exp\left(-\frac{\mathbf{z}_{u2}}{2\sigma'^2}\right)\exp\left(-\frac{\mathbf{z}_{v1}}{2\sigma'^2}\right)}{\left[\exp\left(-\frac{\mathbf{z}_{u1}}{2\sigma'^2}\right) + \exp\left(-\frac{\mathbf{z}_{u2}}{2\sigma'^2}\right)\right]\left[\exp\left(-\frac{\mathbf{z}_{v1}}{2\sigma'^2}\right) + \exp\left(-\frac{\mathbf{z}_{v2}}{2\sigma'^2}\right)\right]}\right|.\end{aligned} \tag{9}$$

Since each exponential term lies in $[0, 1]$, the denominator is bounded above by 4. Letting it be denoted by $\exp(m)$, we apply the Lagrange mean value theorem and derive:

$$\begin{aligned}&|\mathbb{P}(\hat{y}_u = c_1|\mathbf{h}_u) - \mathbb{P}_2(\hat{y}_v = c_1|\mathbf{h}_v)| \\ &= \left|\exp\left(-\frac{(\mathbf{z}_{u1} + \mathbf{z}_{v2})}{2\sigma'^2} - m\right) - \exp\left(-\frac{(\mathbf{z}_{u2} + \mathbf{z}_{v1})}{2\sigma'^2} - m\right)\right| \\ &\leq \frac{1}{2\sigma'^2}\left|-(\mathbf{z}_{u1} + \mathbf{z}_{v2}) + (\mathbf{z}_{u2} + \mathbf{z}_{v1})\right| \\ &= \frac{1}{2\sigma'^2}\left|-\left(\left(\mathbf{h}_u - \boldsymbol{\mu}_1'\right)^2 + \left(\mathbf{h}_v - \boldsymbol{\mu}_2'\right)^2\right) + \left(\left(\mathbf{h}_u - \boldsymbol{\mu}_2'\right)^2 + \left(\mathbf{h}_v - \boldsymbol{\mu}_1'\right)^2\right)\right| \\ &= \frac{1}{\sigma'^2}||\mathbf{h}_u - \mathbf{h}_v|| \cdot ||\boldsymbol{\mu}_1' - \boldsymbol{\mu}_2'|| \\ &= \frac{1}{\sigma'^2}||\mathbf{h}_u - \mathbf{h}_v|| \cdot ||\frac{1+\tau}{2}\boldsymbol{\mu}_1 - \frac{1-\tau}{2}\boldsymbol{\mu}_2 - \frac{1+\tau}{2}\boldsymbol{\mu}_2 + \frac{1-\tau}{2}\boldsymbol{\mu}_1|| \\ &= \frac{1}{\sigma'^2}||\mathbf{h}_u - \mathbf{h}_v|| \cdot ||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2||.\end{aligned} \tag{10}$$

This completes the proof. □

Theorem 1.1 shows that the consistency of predictions for nodes within the same class is affected by both feature similarity and label similarity. In contrast, as pointed out in [3], for traditional GNNs such as SGC [4], the probability difference between predicting two nodes as belonging to class c_1 is bounded by

$$|\mathbb{P}(\hat{y}_u = c_1|\mathbf{h}_u) - \mathbb{P}(\hat{y}_v = c_1|\mathbf{h}_v)| \leq \frac{\rho}{\sqrt{2\pi}\sigma}\left(||\mathbf{h}_u - \mathbf{h}_v|| + \rho\,|e_u - e_v|\right), \tag{11}$$

where the SGC model follows a simplified propagation rule $\hat{\mathbf{Y}} = \mathrm{softmax}(\mathbf{AXW}^k)$, removing non-linear activation functions between layers. Here, $e_i = \frac{|j \in \mathcal{N}(i): y_j = y_i|}{d_i}$ denotes the local homophily ratio of node i. Compared with Eq. 11, GRASP achieves a tighter upper bound on the probability difference, indicating a higher likelihood that nodes from the same class will be classified consistently. This improvement stems from GRASP's ability to construct homophilic and heterophilic subgraphs, where nodes exhibit more uniform local homophily. As a result, the model generates more discriminative and robust embeddings, ultimately leading to improved classification performance. Overall, this theoretical result provides strong support for the core design of GRASP's graph-splitting strategy, highlighting it as a principled and effective approach to addressing heterophily in graph-structured data.

2. EXTRA EXPERIMENTAL RESULTS

2.1. Datasets and Setups

In this section, we conduct experiments to evaluate the performance of GRASP on node classification across various graph benchmarks. Specifically, we consider two types

Table 1: The statistics of the datasets

Dataset	#Nodes	#Train	#Val	#Test	#Edges	#Feat.	#Classes
Cora	2708	1625	541	542	5429	1433	7
Citeseer	3327	1996	665	666	4732	3703	6
Pubmed	19717	11830	3943	3944	44338	500	3
Chameleon	2277	1366	455	456	36101	2325	5
Cornell	183	109	37	37	298	1703	5
Texas	183	109	37	37	309	1703	5
Squirrel	5201	3121	1040	1040	217073	2089	5

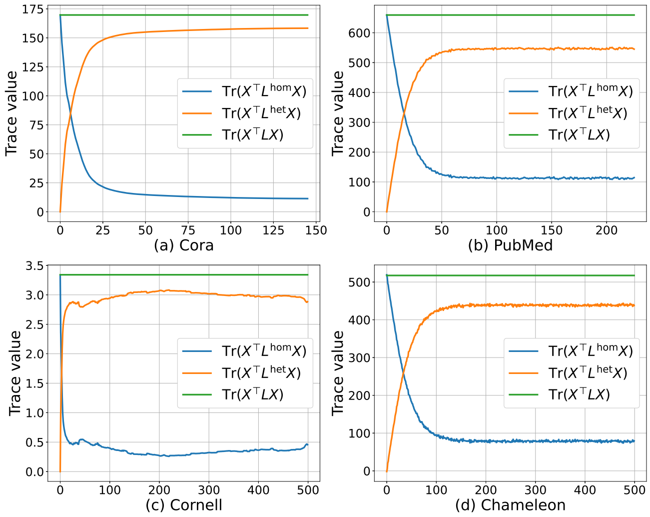

Fig. 1: The behavior of embedding smoothness and roughness regularizers during training on Cora, PubMed, Cornell and Chameleon. The x-axis stands for the training epochs.

of datasets, and the dataset statistics details can be found in Tab. 1.

We report the mean and standard deviation of the classification accuracy over ten independent runs to ensure the robustness of our findings. All experiments are conducted on two NVIDIA GeForce RTX 3090 GPUs, each with 24GB of memory. We employ a random search strategy to identify the best hyperparameters for each dataset. Specifically, we define a search space for key hyperparameters and randomly sample configurations during training. This approach allows us to explore different parameter settings efficiently without introducing human bias.

2.2. Model Analysis

2.2.1. Trace Curve

Fig. 1 shows the behavior of the embedding smoothness $\mathcal{L}_{\text{ES}}$ and roughness $\mathcal{L}_{\text{ER}}$, during the training process on for example datasets Cora, Pubmed, Cornell and Chameleon. As the training progresses, the embeddings on the homophilic subgraphs become smoother, while those on the heterophilic subgraphs become rougher. This observation highlights GRASP's ability to effectively split the original graph into homophilic and heterophilic subgraphs and thus decouple different frequency information, successfully .

Table 2: Ablation study. We highlight the best results in bold.

Ablation	Cora	Pubmed	Cornell	Chameleon
w/o rel	90.01	89.30	93.42	73.67
$\mathcal{L}_{\text{F}}$-hom	89.11	91.36	85.94	69.54
$\mathcal{L}_{\text{F}}$-het	84.31	90.89	90.79	69.92
w/o $\mathcal{L}_{\text{ES}}$	85.06	88.10	83.78	65.29
w/o $\mathcal{L}_{\text{ER}}$	87.82	90.03	84.68	68.82
GRASP	**90.77**	**91.40**	**93.82**	**73.78**

2.2.2. Parameters Sensitive Analysis

In this subsection, we conduct a sensitivity analysis on several key hyper-parameters in our model, including weight decay, Frobenius norm weight (α), smoothness and roughness regularizers for the homophilic and heterophilic subgraphs (β and γ), the balance between low- and high-pass filters (τ), hidden dimension size, learning rate of GNN (η), and the learning rate of the adjacency matrix ($\eta^{'}$). To ensure fair evaluation, we fix all other hyper-parameters to their optimal values when tuning a specific parameter, and the experiments are performed on the two homophilic and two heterophilic graphs: Cora, Pubmed, Cornell and Squirrel. Specifically, weight decay is varied from 5e-6 to 5e-4, α, β, and γ from 5e-6 to 5e-2, τ from 0 to 1, hidden size from 16 to 128, and both η and $\eta^{'}$ from 5e-4 to 5e-2.

The experimental results are illustrated in Fig. 2. To better understand the impact of key components in GRASP, we conduct a brief analysis on the core hyperparameters. The Frobenius norm weight α leads to the best result at a moderate value, while overly small or large values harm performance, as they either under-regularize or over-constrain the graph structure learning. Both β and γ show clear trends: increasing their values improves accuracy up to a point, with the best performance observed, validating the importance of smoothness and roughness regularization. For the mixing ratio between low- and high-pass filters τ, the result indicates that balancing homophilic and heterophilic information leads to better representations. Too much emphasis on either type of information results in degraded performance. Overall, these results highlight the significance of careful hyper-parameter tuning and confirm that each component—regularization, smoothness, and frequency balance—plays a vital role in GRASP's capability of learning meaningful embeddings.

3. REFERENCES

[1] Yash Deshpande, Subhabrata Sen, Andrea Montanari, and Elchanan Mossel, "Contextual stochastic block mod-

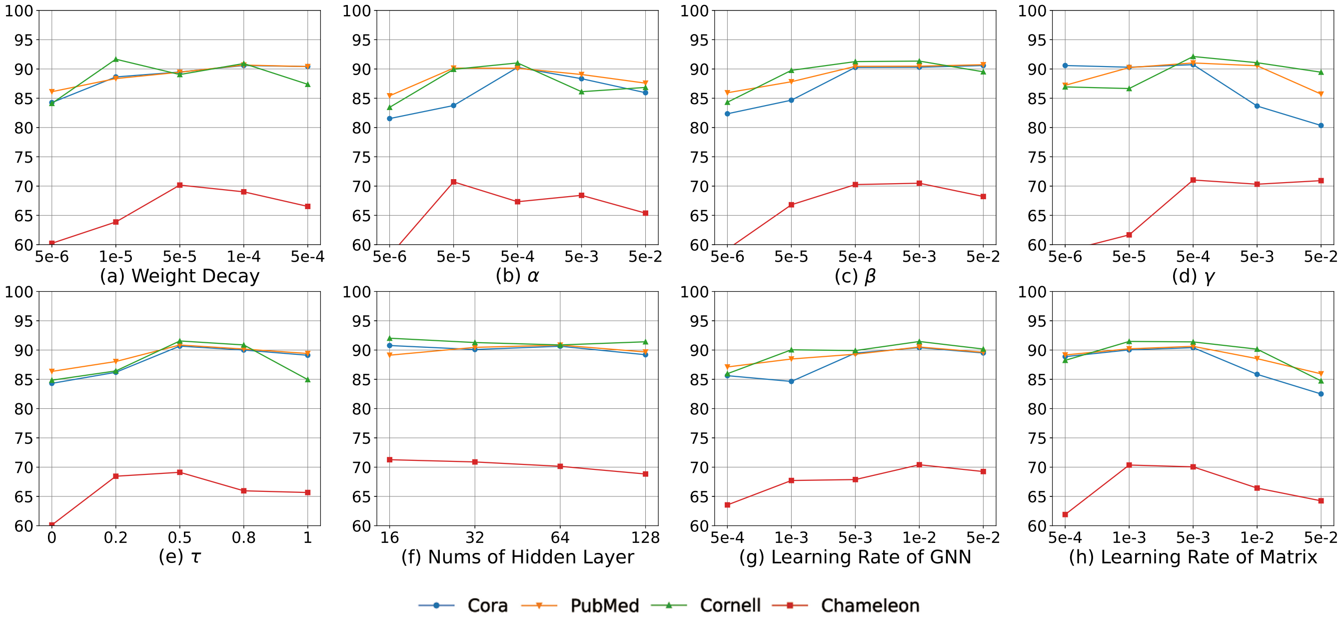

Fig. 2: Parameters sensitive analysis on Cora, Pubmed, Cornell and Chameleon.

els," in *Advances in Neural Information Processing Systems*, 2018.

[2] Xinyi Wu, Zhengdao Chen, William Wang, and Ali Jadbabaie, "A non-asymptotic analysis of oversmoothing in graph neural networks," *arXiv preprint arXiv:2212.10701*, 2022.

[3] Haitao Mao, Zhikai Chen, Wei Jin, Haoyu Han, Yao Ma, Tong Zhao, Neil Shah, and Jiliang Tang, "Demystifying structural disparity in graph neural networks: Can one size fit all?," in *Proceedings of the 37th Conference on Neural Information Processing Systems*, 2023, pp. 37013–37026.

[4] Felix Wu, Amauri Souza, Tianyi Zhang, Christopher Fifty, Tao Yu, and Kilian Q. Weinberger, "Simplifying graph convolutional networks," in *Proceedings of the 36th International Conference on Machine Learning*, 2019, vol. 97, pp. 6861–6871.