ArnoldiGCL: Graph Contrastive Learning via Learnable Arnoldi-Based Guided Spectral Chebyshev Polynomial Filters –Supplementary Materials

MUSTAFA COŞKUN, ABDELKADER BAGGAG AND MEHMET KOYUTÜRK

1. COMPARISON OF ARNOLDIGCL WITH VANDERMONDEGCL

To validate the theoretical findings presented in the main text, we compare our proposed method, `ArnoldiGCL`, against direct Vandermonde matrix inversion. Specifically, we solve the system $\boldsymbol{h} = \boldsymbol{V}\boldsymbol{a}$, which is expected to yield inaccurate a_k coefficients due to the ill-conditioned nature of the Vandermonde matrix—characterized by an exponentially bounded condition number.

For this purpose, we directly utilize `Python`'s `np.linalg.solve(V, y)` function to compute the coefficients. We refer to this approach as `VandermondeGCL` for naming consistency within the context of GCL. More specifically, we analyze `VandermondeGCL` in terms of its ability to fit complex filters and its performance in self-supervised GCL-based node classification.

Effect of Ill-fitting on GCL. To further validate the impact of ill-fitting to any complex filter in the context of Graph Contrastive Learning, we replicate the same experiments conducted for `ArnoldiGCL` using `VandermondeGCL`. Specifically, we evaluate both methods on the `Texas`, `Cornell`, and `Wisconsin` datasets, while keeping the dropout rate and dprate fixed at 0.5 for both algorithms.

The results of this analysis are presented in Figure S1. As observed, `VandermondeGCL` significantly degrades the performance of GCL, emphasizing the critical role of accurate coefficient learning in ensuring effective contrastive representations.

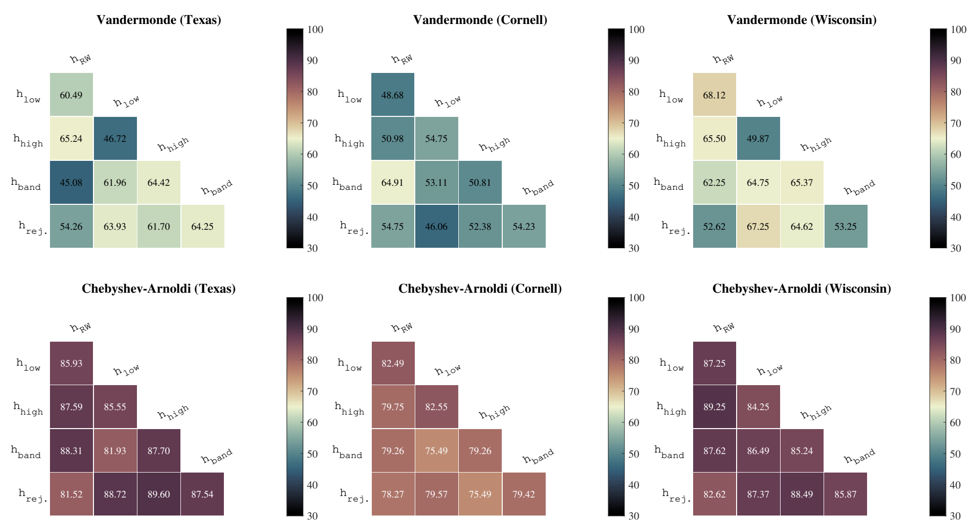

Fig. S1. `VandermondeGCL` Versus `ArnoldiGCL` for small heterophilic datasets

2. VISUALIZATION OF NODE EMBEDDINGS

To further investigate the effect of the use of complex filters on the embeddings learned by the `ArnoldiGCL` algorithm, we perform a qualitative analysis using t-SNE visualizations. For this purpose, we select two datasets from each group of benchmark datasets: `Cora` and `Citeseer` representing homophilic graphs, `Texas` and `Wisconsin` for small-scale heterophilic graphs, and `Minesweeper` and `Roman-Empire` for large-scale heterophilic graphs.

In these experiments, we set both the `dropout` and `droprate` hyperparameters to 0.5. We evaluate the following combinations of *spectral filters* within `ArnoldiGCL`: `Random Walk (G-1) - Band Rejection`, `Low Pass - High Pass`, `Low Pass - Band Pass`, and `Band Pass - Band Rejection`. The resulting t-SNE visualizations of the node embeddings produced by these filter combinations are presented in Figures S2–**??**.

From these experiments, we observe that the choice of filter combinations significantly influences the structure of the embeddings. Specifically, as seen in the figures, filter pairs that yield higher classification accuracy tend to produce more clearly separated clusters in the t-SNE plots. In contrast, combinations associated with lower classification performance often result in more overlapped and less distinguishable t-SNE embeddings.

For homophilic networks, we observe that GCLs that utilize at least one homophilic network provides better separation than GCLS that utilize no homophilic filters and this translates into better classification accuracy. This observation suggests that inclusion of non-homophilic filters may be creating undue confusion on homophilic networks. However, as seen for the Citeseer network, the use of the high-pass filter along with the random walk filter improves the dispersion of the nodes across the lower-dimensional space, resulting in improved accuracy over models that utilize band-reject or band-pass filter in addition to a homophilic (random walk or band-pass) filters.

For small-scale heterophilic networks, the embeddings exhibit different behavior as compared to the homophilic networks. The embeddings that perform well in classification appear to have correlated dimensions for the lower-dimensional embeddings, exhibiting strong linear correlation between the two dimensions of the tSNE embedding used for visualization. For both the Texas and Wisconsin datasets, the embeddings that perform well in classification utilize at least one banded filter (band-pass or band-reject) and for both datasets the GCL that combines a band-pass filter with a low-pass filter performs best. This observation suggests that, while these networks are heterophilic, nodes that are in close and medium proximity of other nodes are informative of each others' labels, while distant nodes are not very informative.

The visualization of the embeddings for the large-scale heterophilic networks demonstrates that these are the most difficult instances for node classification. For the Minesweeper dataset, it is difficult to discern the differences between the embeddings computed using different visualizations from t-SNE visualizations - however, it is interesting to note that all filter combinations appear to induce similar node clusters. The combination of random walk and band-reject filters appears to result in the tightest clusters and deliver best classification accuracy, but the underlying reasons for this behavior require further investigation. For the Roman-Empire dataset, some labels appear to induce more coherent clusters, and the combination of a low-pass and high-pass filter appears to induce coherent clusters for more than one node class, resulting in higher classification accuracy. Taken together, these results suggests that the ability to incorporate complex filters holds significant potential in capturing the intricate relationships between node classes and network topology.

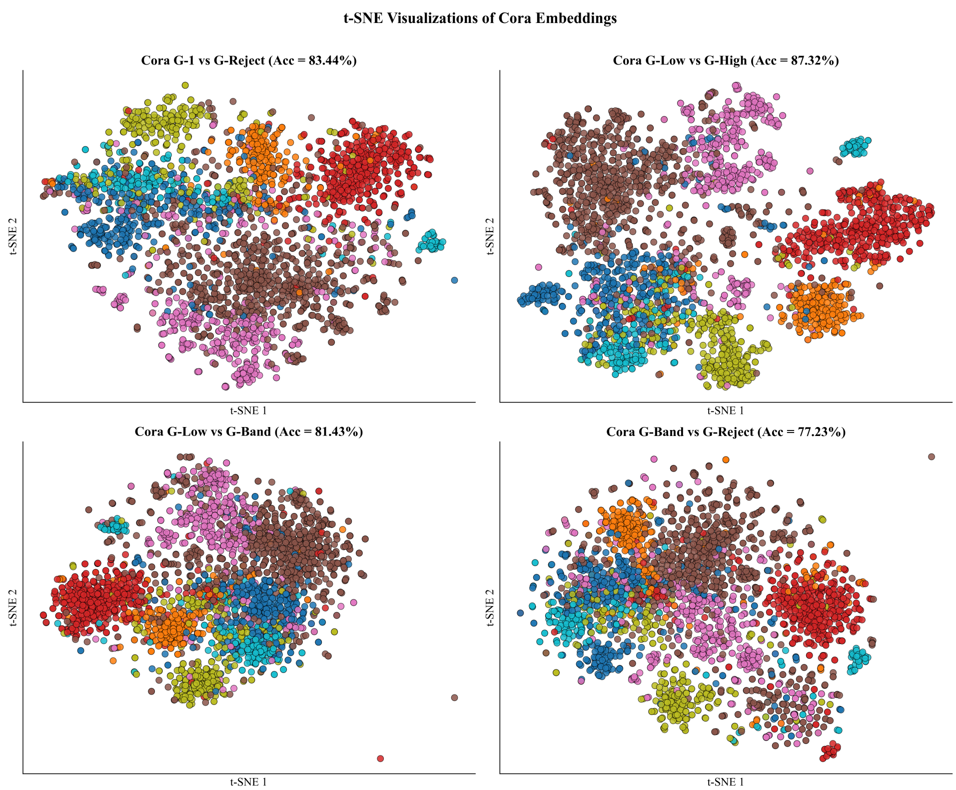

Fig. S2. Node embeddings learned by `ArnoldiGCL` using different pairs of filter combinations on the Cora dataset.

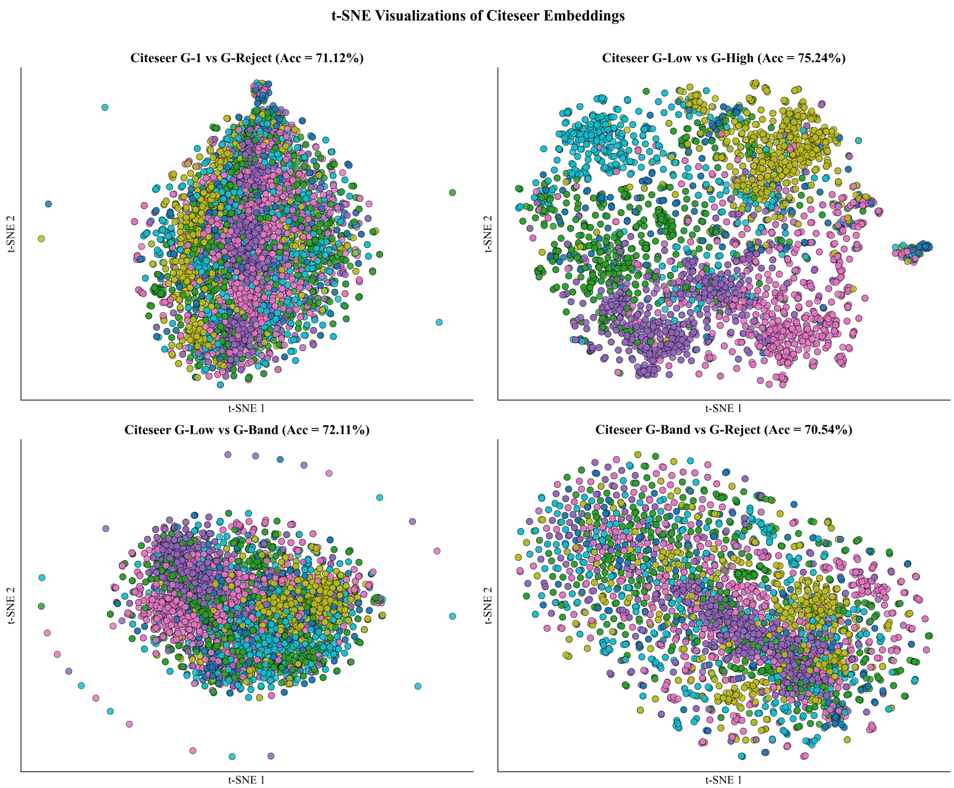

Fig. S3. Node embeddings learned by `ArnoldiGCL` using different pairs of filter combinations on the CITESEER dataset.

Fig. S4. Node embeddings learned by `ArnoldiGCL` using different pairs of filter combinations on the Texas dataset.

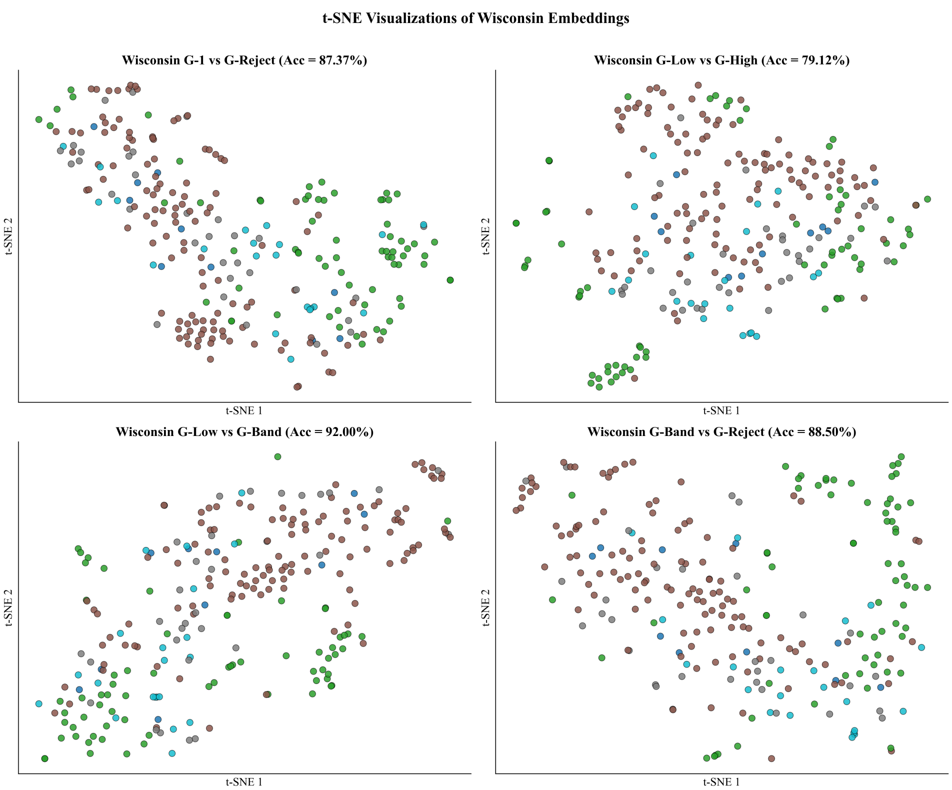

Fig. S5. Node embeddings learned by `ArnoldiGCL` using different pairs of filter combinations on the Wisconsin dataset.

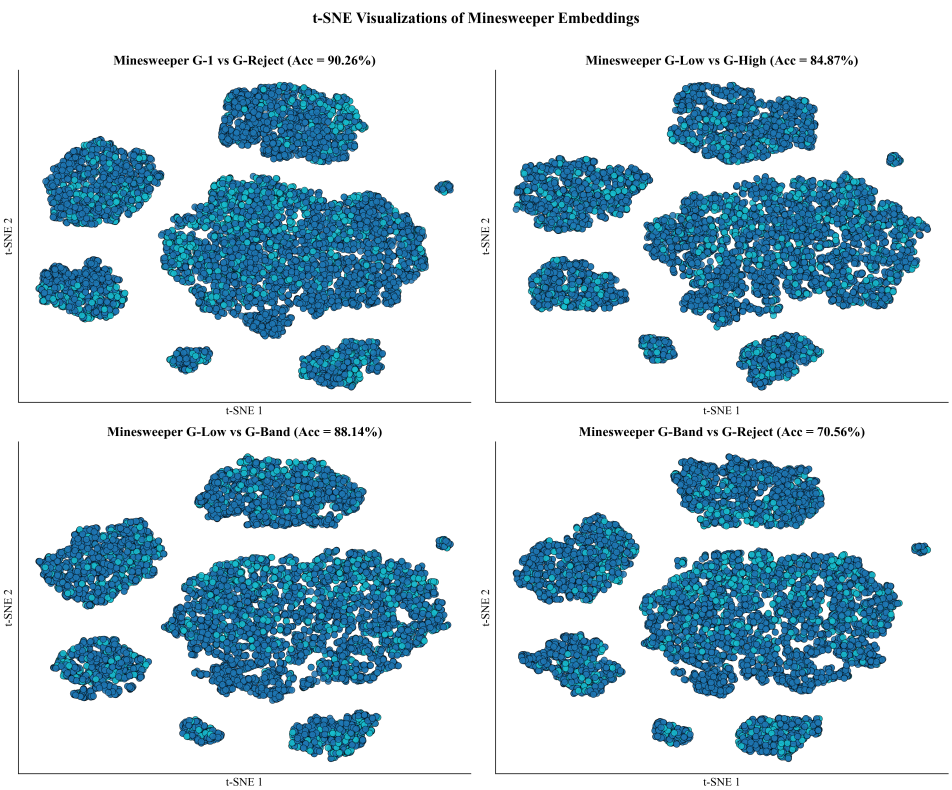

Fig. S6. Node embeddings learned by `ArnoldiGCL` using different pairs of filter combinations on the MINESWEEPER dataset.