Characterization of Simplicial Complexes by Counting Simplets Beyond Four Nodes: Online Appendix

1 CHARACTERIZATION OF NODES IN A SIMPLICIAL COMPLEX

Simplets and SC3 can also be used to characterize nodes, and the results can be used as input for node-level tasks. Specifically, we can characterize a node in an SC by the absolute count of each simplet in its egonet. For an SC $\mathcal{K} = (V, E)$, the *egonet* of a node $v \in V$ is the subcomplex induced by v and its 1-hop neighbors, i.e., $\mathcal{K}[V_v]$ where $V_v = \bigcup_{\sigma \in E \text{ s.t. } v \in \sigma} \sigma$. As a result of characterization, each node is represented as a integer vector whose dimensionality is the same as the number of simplets (i.e., s_k).

2 APPLICATION: NODE CLASSIFICATION

We present an experiment on node classification.

Data: We used the 12 datasets from the following domains:

- **coauthership**: cD, cMG, cMH
- **school**: chs, cps
- **email**: eEu, eEn
- **tags**: taau, taso
- **threads**: thau, thms, thso

For abbreviations, refer to Table 4 in the main paper. For each dataset, we use 100 nodes uniformly at random among those whose egonet has at least 10 nodes and at most 10^4 nodes. We use the domain of each dataset as the class of each node in the dataset. As a result, we use 1,200 nodes and the following five classes: coauthership, school, email, tags, and threads.

Experimental Protocol: We split the nodes randomly into a training set (80%) and a test set (20%). For each node, we repeated SC3 five times and averaged the counts of simplets in its egonet, which were used as input features, as described in Section 1. We used a gradient boosting method [1, 2] implemented in the Python scikit-learn library as a classifier.

Competitors: We compared SC3 with two simple baselines: (1) majority selection (i.e., yielding always a majority class as output), and (2) random guessing.

Results: The node-classification results are presented in Table 1. We use f1-micro scores as the accuracy measure. Our simplet-based approach significantly outperformed both baselines. Interestingly, its accuracy was slightly higher when $k = 4$ than when $k = 5$.

Table 1: F1-micro Scores on Node Classification.

k	simplet based method	majority selection	random guessing
4	0.88	0.25	0.2
5	0.85	0.25	0.2

REFERENCES

[1] Leo Breiman. 1997. *Arcing the edge.* Technical Report. Citeseer.

[2] Jerome H Friedman. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics* (2001), 1189–1232.