Exploring a link between network topology and active learning

Michael Hopwood*†, Phuong Pho*, Alexander V. Mantzaris*
*Department of Statistics and Data Science, University of Central Florida, Orlando, FL 32816 USA
†Sandia National Laboratories, Albuquerque, NM 87123 USA

***Abstract*—In many online networking platforms whether they are online social networks or academic citation networks, community label memberships are applied to the nodes to generalize overlaps based upon common features or associations. The relatively recent methods in graph convolutional neural networks has provided new tools to infer community labels of nodes but they still depend upon a labeled dataset to be provided. Obtaining these labels can be costly and methods to reduce the required number of labels needed can speed up the process and reduce costs. This study explores different strategies for selecting nodes to be used as training data showing which strategies work better or worse and on different percentages of the network's nodes. An unsupervised approach of deciding the best active learning sampling direction (i.e. ascending or descending selection of nodes in terms of importance) procedure is derived by fundamental network properties. The conclusion is supported on both simulated and real data.**

***Index Terms*—network science, social networks, machine learning, graph neural networks, active learning**

I. Introduction

The study of networks [1] shows that they are ubiquitous in nature and commerce. Many complex processes that are studied [2] have shown that behaviors are defined or closely dependent on their network structure. Simplifying network nodes into a smaller set of labels, commonly through community detection where the connectivity of the nodes in the network directs the community memberships, is a common task in this space. This labelling process assists in the effort to simplify the node set by generalizing them with respect to aggregated behaviors across allocated groups. The principle underlying the ability to group nodes together in this fashion relies on homophily [3]. Examples of this are found in the work of [4] which studies how social network connections created from friendships or interests can drive political engagements differently.

Labels of nodes can be inferred using standard classification methods such as logistic regression, which are predominantly reliant on the node feature information, after a training phase. However, these methods do not consider the supplementary node connectivity information. Additionally, community detection algorithms (i.e. Louvain [5]) take into account node connectivity information but not node feature information. Graph neural networks (GNN) combine both information into a framework for inference (i.e. label prediction). The Simple Graph Convolutional neural network (SGC) [6] (see Methodology section for more details) simplifies GNNs to a logistic-regression-like formulation while maintaining the node connectivity information. The computational efficiency of this model allows the practical experimentation done in this study.

In many cases, labeled data is limited and costly to produce. The field of active learning focuses on ordering the available labeled data prior to the training process for the purpose of strategically showing the model more informative nodes earlier, allowing it to generalize with less data while maintaining a similar (or superior) performance [7]. This paper focuses on the application of active learning to graph neural networks (GNN) by utilizing available node ranking algorithms such as node connectivity densities (i.e. degree), pagerank [8], and voterank [9]. Similar to [10], this work experiments with bidirectional sampling (i.e. ascending and descending) of these algorithms' rankings.

Node classification task across four real graph datasets are optimized using the six node selection processes (i.e. ascending & descending selection along 3 node importance evaluators) to study the correlation between the superior sampling process and network topology. Results show that the sampling direction (i.e. ascending vs descending selection of samples with respect to their importance rankings) is dependent on network topology. The results are empirically reverse engineered using an unsupervised process to allow the prediction future applications to derive the best sampling method as opposed to the brute force experimentation provided in this study. Generally, networks with sparse topologies are better performant in node classification tasks when the active learning process uses a descending node selection; conversely, dense networks prefer ascending node selection.

Section II delineates the methodology, including descriptions of the data, sampling methods, and the employed graph convolutional neural network. Section III shows the results along with a discussion of the results. Lastly, section IV describes the final takeaways and some potential future work.

II. Methodology

A. Data

An attributed graph $G = (X, A, y)$ is represented by three components: an adjacency matrix $A \in \mathbb{R}^{N \times N}$, a feature matrix $X \in \mathbb{R}^{N \times D}$, and a node label vector $y \in \mathbb{R}^{N}$. Real datasets were gathered from online resources [11]–[14]. Synthetic attributed graph data sets were generated to imitate scale-free (right-skew degree distribution) networks which are

Dataset	Ref.	#Nodes #Edges #Classes	Description
Cora	[11]	2708 5278 7	Scientific publications (nodes), defined by a binary vector indicating the presence of words in the paper (features), connected in a paper citation web (edges) and categorized by topic (labels).
Citeseer	[12]	3327 4614 6	Scientific publications (nodes), defined by a binary vector indicating the presence of words in the paper (features), connected in a paper citation web (edges) and categorized by topic (labels).
Pubmed	[13]	19717 44325 3	Diabetes-focused scientific publications (nodes), defined by a binary vector indicating the presence of words in the paper (features), connected in a paper citation web (edges) and categorized by topic (labels).
Amazon-Photo	[14]	7650 143663 8	Photos sold at Amazon (nodes), defined by a bag-of-words encoded vector of the product's reviews, connected in groups of products which are frequently bought together (edges), and grouped into product categories.

commonly found in practice (Figures 1 and 2) by using the Barabási–Albert preferential attachment model [15].

Each synthetic attributed graph contains three communities clusters (subgraphs) with 100 nodes per community. Each subgraph is generated following Barabási–Albert preferential attachment model. On each subgraph, we collect a subset of nodes (using weighted random sampling proportional to node degree distribution) and then assign random edges between pairs of subsets of nodes. Hyperparameters (number of preferential attachment for Barabási–Albert model, probability of random edges) are then established to control the connectivity between subgraphs (seen in Figures 1 and 2). The node feature matrix is generated by first creating a set of three isotropic Gaussian clusters (100 observations per each cluster) in a two dimensional feature space and then assigning these observations as node features. We control the amount of overlap between three clusters by adjusting two cluster hyperparameters (the distances between cluster centers and the within cluster standard deviation).

B. Sampling methods

Two procedures of sampling are considered in this study, namely descending and ascending. In descending sampling, training instances are selected by gradually acquiring from the most important nodes to the least important ones. On the contrary, ascending sampling gradually selects training samples starting from the least important nodes to the most important ones.

Three different criteria are used to evaluate a node's importance (centrality) for sampling orders. In degree sampling, we acquire nodes for training based on their corresponding number of directly connected neighbours (i.e node's degree). The PageRank algorithm [8] derives a web page (node)'s rank by accumulate its incoming neighbors' ranks proportionally to their total number of outgoing connections. The resulting ranking represents the relative importance of pages in the network. In this study, we apply PageRank to rank all the nodes in our graphs and then sample them based on their rankings. Lastly, the VoteRank algorithm [9] iteratively selects a set of important nodes called spreaders using voting scores given by the neighboring nodes. Once a node is selected as spreader, it is excluded from next round of voting and its direct neighbors' voting abilities are also reduced. In this study, we employ VoteRank to all nodes in the graph (by setting the number of spreaders as the total number of nodes) and then sample them based on their rankings.

C. Simple Graph Convolution (SGC)

SGC [16] is a simplified GNN model developed from GCN [17] by removing non-linear activation functions between hidden layers and reparametrizing successive layers into one single layer. This simplification reduces superfluous complexity of GCN while retains superb performance on many downstream tasks. The work of [18] illustrates SGC's expressive power in node classification task and proposes a flexible regularization methodology to reduce the number of parameters and highlight a sparse set of important features.

In this section, we briefly present the original SGC. An attributed graph data set contains a graph $G = (V; \mathbf{A})$ and a feature matrix $X \in \mathbb{R}^{N \times D}$. The graph G composes of $V = (v_1, v_2, ..., v_N)$ is a set of N nodes (vertices) and $\mathbf{A} \in R^{N \times N}$ is the adjacency matrix where each element a_{ij} represents an edge between node v_i and v_j ($a_{ij} = 0$ if v_i and v_j are disconnected). We define the degree matrix $\mathbf{D} = \text{diag}(d_1, d_2, ..., d_N)$ as a diagonal matrix whose off-diagonal elements are zero and each diagonal element d_i capture the degree of node v_i and $d_i = \sum_j a_{ij}$. Each row x_i of the feature matrix $X \in \mathbb{R}^{N \times D}$ is the feature vector measured on each node of the graph. Each node i receives a label from C classes and hence can be coded as one hot vector $y_i \in \{0, 1\}^C$.

The GCNs and SGC add self-loops and normalize the adjacency matrix to get the matrix $\mathbf{S}$:

$$\mathbf{S} = \tilde{\mathbf{D}}^{-\frac{1}{2}} \tilde{\mathbf{A}} \tilde{\mathbf{D}}^{-\frac{1}{2}} \tag{1}$$

where $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}$ and $\tilde{\mathbf{D}} = \text{diag}(\tilde{\mathbf{A}})$. This normalization allows successive powers of the matrix to not influence the overall size the projections. The SGC removes non-linear

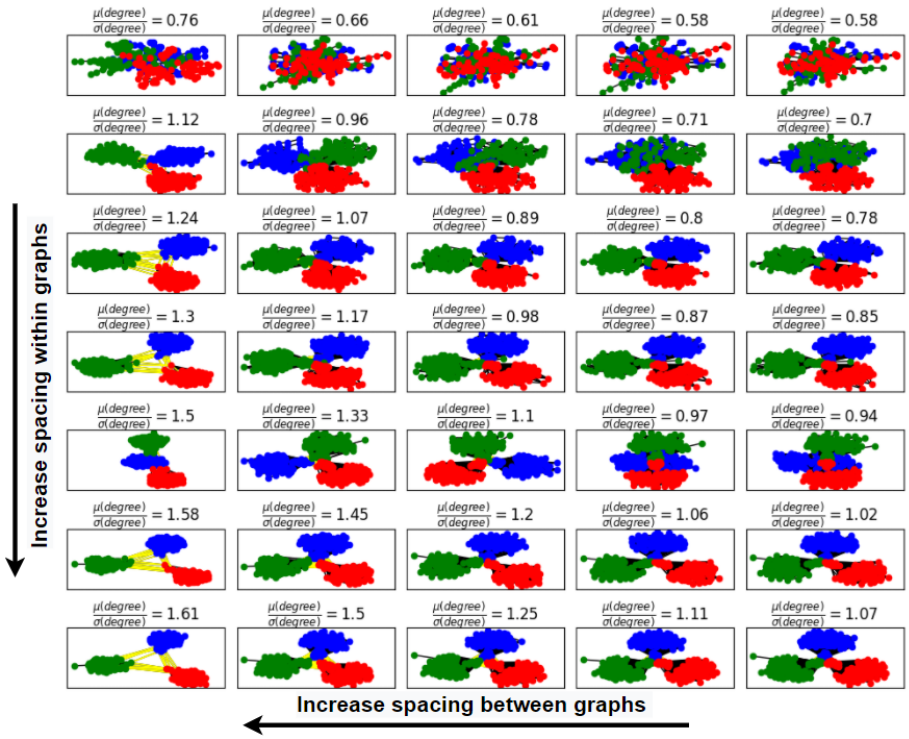

Fig. 1. Network visualizations for the 35 generated simulations, each with 3 communities (colored). Traversing along the y-axis shows how these networks topologies change when varying the distance within a communities. Traversing along the x-axis shows how the network topologies changes when varying distance between communities.

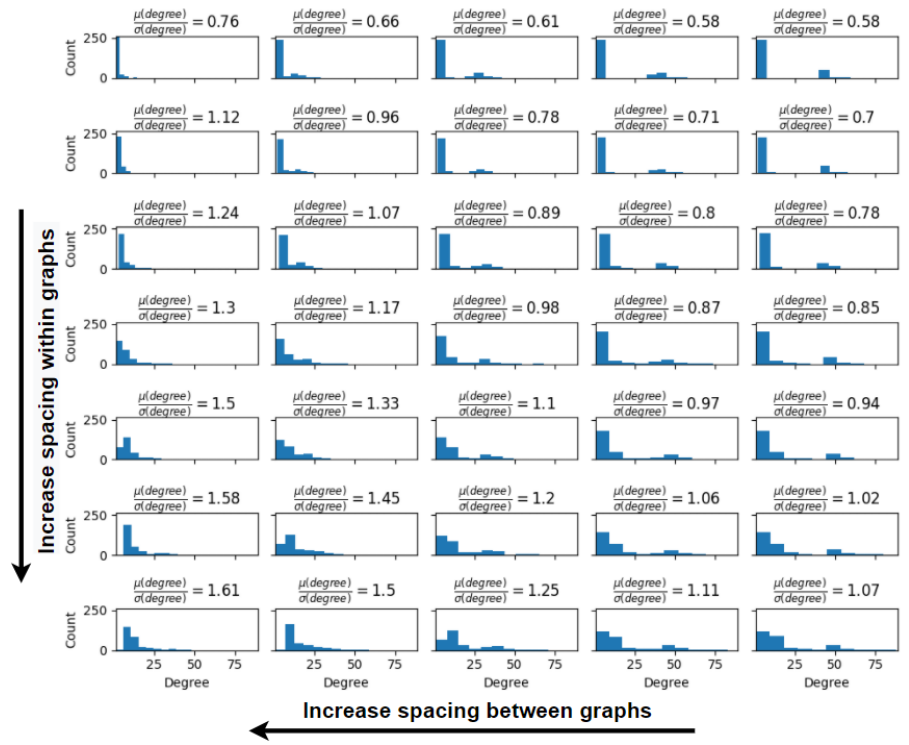

Fig. 2. Degree distributions for the 35 generated simulations show that all settings create a relatively scale-free network. Traversing along the y-axis shows how these networks topologies change when varying the distance within a communities. Traversing along the x-axis shows how the network topologies changes when varying distance between communities.

transformation from the k^{th}-layer of the GCN resulting in a linear model of the form:

$$\hat{\mathbf{Y}} = \text{softmax}(\mathbf{S} \dots \mathbf{S}\mathbf{S}\mathbf{X}\mathbf{\Theta}^{(1)}\mathbf{\Theta}^{(2)} \dots \mathbf{\Theta}^{(K)}). \quad (2)$$

The SGC classifier is then achieved by collapsing the repetitive multiplication of matrix $\mathbf{S}$ into the k^{th} power matrix $\mathbf{S}^K$ and reparameterizing the successive weight matrices as $\mathbf{\Theta} = \mathbf{\Theta}^{(1)}\mathbf{\Theta}^{(2)} \dots \mathbf{\Theta}^{(K)}$:

$$\hat{\mathbf{Y}} = \text{softmax}(\mathbf{S}^K\mathbf{X}\mathbf{\Theta}). \quad (3)$$

The parameter k corresponds to the number of 'hops' which is the number of edge traversals in the network adjacency matrix $\mathbf{S}$. k can be thought of as accumulating information from a certain number of hops away from a node (as described visually in [16]). If $k = 0$ the methodology becomes equivalent to a logistic regression application which is known to be scalable to large datasets. Since the SGC introduces the matrix $\mathbf{S}$ as linear operation the same scalability applies. The weight matrix $\mathbf{\Theta}$ is trained by minimizing the cross entropy loss:

$$\mathcal{L} = \sum_{l \in \mathcal{Y}_L} \sum_{c \in C} Y_{lc} \ln \hat{Y}_{lc} \quad (4)$$

where $\mathcal{Y}_L$ is a collection of labeled nodes.

D. Evaluation of Network Topology

The network topology was evaluated using the coefficient of variation of the node's degree distribution.

$$CV_d = \frac{\mu_d}{\sigma_d} \quad (5)$$

where $\mu_d = \frac{1}{N}\sum_{i=1}^{N} d_i$ is the average degree and $\sigma_d = \frac{1}{N-1}\sum_{i=1}^{N}(d_i - \mu)^2$ is the standard deviation of degree.

A low value of CV_d occurs for networks which have high variation in their degree distributions compared to the mean degree. It indicates that important hubs (nodes) are highly connected to other nodes. On the contrary, a high value of CV_d results from relatively low variation in degree distribution compared to the mean degree where important nodes tends to be less popular.

The feature information was evaluated using the coefficient of variation of the node-to-node feature distances. For example, node 1 is defined as a vector of distances between its feature vector and all other nodes' feature vectors. This description allows for an evaluation regarding a node's centrality in the feature space.

III. Results

In this section, the correlation of the optimal sampling direction for node classification task with network topology is captured in simulations and real data. The results show that no sampling method (i.e. degree, Pagerank, Voterank) is uniformly superior in terms of accuracy. However, independent of the ascending/descending, we see across the board a higher number of cases where the more complicated sampling procedures (i.e. Pagerank/Voterank) outperform Degree. While we see an increase in performance, there is a trade-off with computation time; nodes degree distribution can be computed swiftly while Pagerank and VoteRank require complex evaluation and hence, be more computationally expensive.

Dataset	CV_d	Optimal sampling direction
Cora	0.75	Descending
Citeseer	0.82	Descending
Pubmed	0.6	Ascending
Amazon-Photo	0.69	Ascending

TABLE I
OPTIMAL SAMPLING RESULTS ON REAL DATASETS

Dataset information including degree coefficient of variation and optimal sampling direction, as derived through a grid search, can be found in Table I. The descending and ascending optimal sampling directions are cleanly partitioned in the CV_d space. A numerical boundary would be useful to allow users to calculated CV_d and perform the active learning procedure without having to experiment through grid search, as done in this paper. From the results (Table I), it is hard to pinpoint an exact threshold other than that it should likely be somewhere between $CV_d = 0.69$ and $CV_d = 0.75$. Therefore, simulations are conducted to resolve a finer resolution. Figure 3 shows a contour plot containing the density of the CV_d distribution for ascending and descending sampling procedures. A partition is found near $CV_d = 0.82$, which is slightly higher than our estimated window, which is likely caused by discrepancies between the real and simulated data. Generally speaking, however, these visualized distributions show significant levels of partitioning on the vertical axis (CV_d). In fact, a one-sided t-test results concludes significance (pval $= 6.7 \times 10^{-41}$).

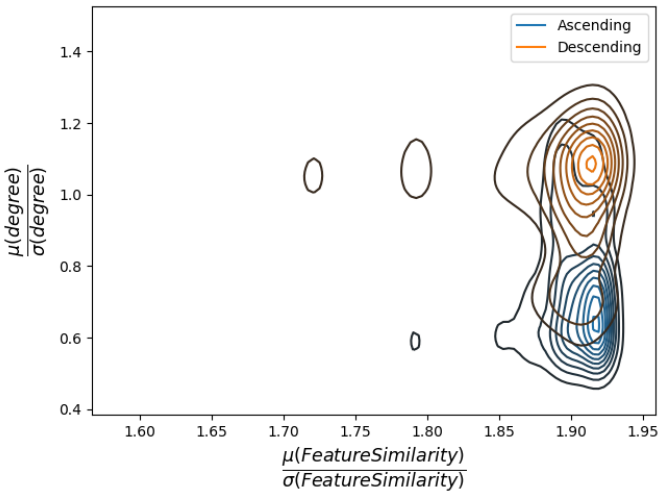

Fig. 3. Simulations report a density of preferred (higher accuracy) sampling direction as a function of network topology (y-axis) and feature similarity (x-axis) shows that the sampling direction is dependent on the network topology.

In high CV_d graphs (i.e Cora and Citeseer), all three descending methods almost uniformly do better than the ascending methods across training sizes (Figure 4). Apparent performance improvements are made in terms of accuracy,

especially at low train sizes (from $s = 0.1$ to $s = 0.5$). As the training size gets closer to utilizing the full training dataset ($s = 1.0$), sampling approaches are less selective because, by definition, they are using more and more data each iteration. The dominance of descending sampling in these graphs might be explained by the fact that important (central) papers of certain disciplines are usually cited by many papers in that same discipline. Consequently, the most important nodes contains crucial information about the class label and are commonly referenced by papers within its discipline so they are beneficial for node classification task.

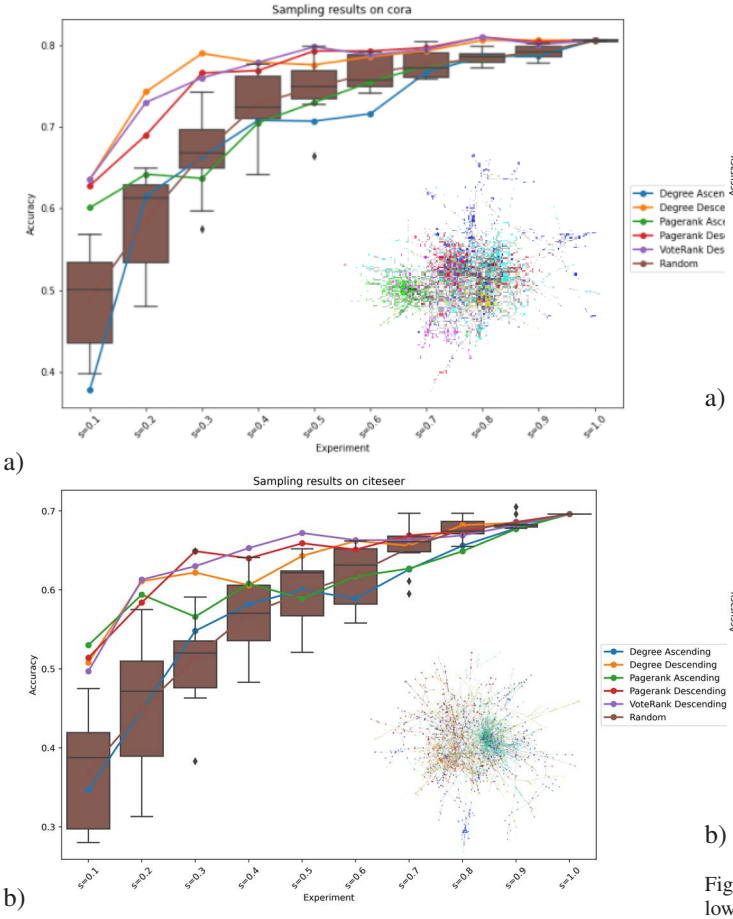

Fig. 4. High CV_d networks (Cora and citeseer) graphs have higher accuracies when sampling nodes from the highest score to lowest score (i.e. 'descending' methods), showing the effectiveness of the node ranking algorithms on a node classification task.

Alternatively, we observe an opposite trend in low CV_d graphs (Pubmed and amazon-photo), where ascending samplings prevail. Pubmed citation graph contains publications about a specified domain and hence has a smaller scope compared with other citation data sets like Cora and Citeseer (Figure 5A). Important (central) papers across classes might cite each other due to the close nature of their categories. Therefore, important nodes contains a less differentiating factor for classification tasks. On the other hand, less important nodes might contain unique characteristics of the class and render useful information for node classification task. Amazon-photo graph exhibits closely connected clusters with relatively low inter-cluster connectivity (Figure 5B). Popular photos from different categories might possess similar features (in term of reviews since they receive generally positive compliments). Hence, sampling popular nodes is less desirable for classification since their representations are indiscriminative. Less popular photos might contain more defined characteristics of its corresponding category. Therefore, lower score nodes may contain higher information about the community and hence be more beneficial for node classification.

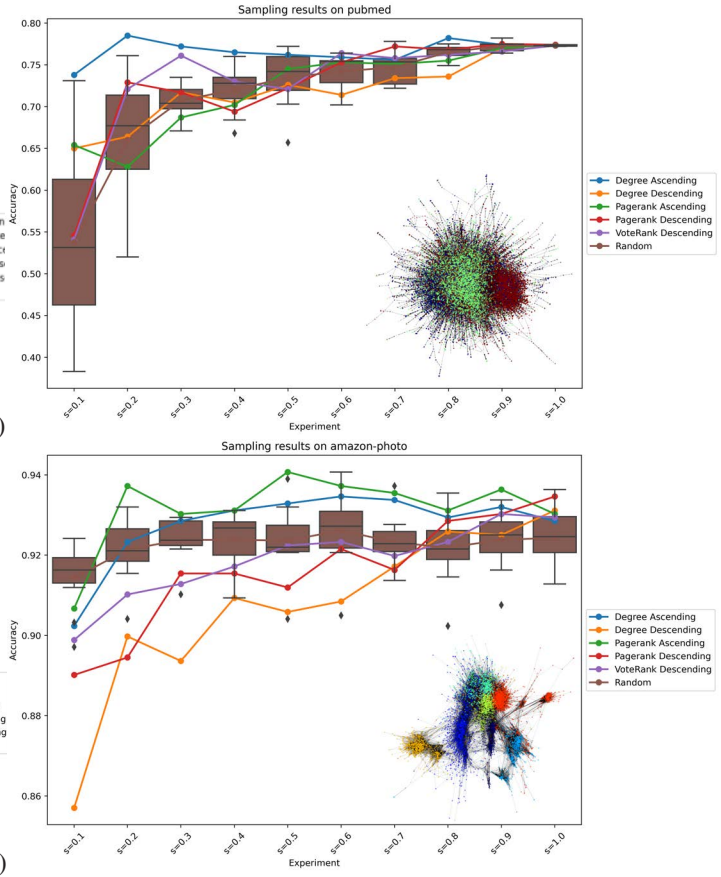

Fig. 5. Low CV_d networks (Pubmed and amazon-photo) graphs have lower accuracies when sampling nodes from the lowest to highest score (i.e. 'ascending' methods), showing the ranking algorithms are inversely beneficial to the node classification task.

IV. Conclusions

Participants in networking platforms continue to upload more data onto these platforms such as in academic literature [19] and social networking [20] (being part of the *always-on* generation [21], and efficient protocols [22]). It becomes a question of efficiency of whether a subset of the nodes can be sampled to provide information about other nodes with unknown membership labels, and can be useful for e-health [23]. The study here conducted on a set of networks covering different information sources show that the best indicator for whether nodes should be sampled in terms of ascending or descending centrality is based upon the coefficient of

variation of the degree of the nodes. Intuitively this can be understood as being related to the sparseness of the network topology. An implication of this is that when attempting to infer labels of network participants, in an active learning paradigm, understanding the general degree distribution for communities can determine whether the sampling should be done in the ascending or descending direction. Practitioners can use the general rule of thumb ($CV_d > 0.8$ should use descending, otherwise ascending sampling direction) to avoid the computational burden of computing grid searches.

Future work will entail applications in professional networking sites (ie LinkedIn) to improve the ability to adopt a community of followers of a certain label, or finding the best connections to develop a new affiliation label. These actions can help navigate the labor market [24] for opportunities, or to plan promotions into new markets.

ACKNOWLEDGMENT

This material is supported by the U.S. Department of Energy's Office of Energy Efficiency and Renewable Energy - Solar Energy Technologies Office (as part of the Durable Modules Consortium (DuraMAT), an Energy Materials Network Consortium). Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration. The views expressed in the article do not necessarily represent the views of the U.S. Department of Energy or the United States Government. SAND Report No. SAND2021-8912 C.

REFERENCES

[1] M. Newman, *Networks*. Oxford university press, 2018.

[2] E. Estrada, *The structure of complex networks: theory and applications*. Oxford University Press, 2012.

[3] M. McPherson, L. Smith-Lovin, and J. M. Cook, "Birds of a feather: Homophily in social networks," *Annual review of sociology*, vol. 27, no. 1, pp. 415–444, 2001.

[4] J. Kahne and B. Bowyer, "The political significance of social media activity and social networks," *Political Communication*, vol. 35, no. 3, pp. 470–493, 2018.

[5] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of statistical mechanics: theory and experiment*, vol. 2008, no. 10, p. P10008, 2008.

[6] F. Wu, T. Zhang, A. H. d. Souza Jr, C. Fifty, T. Yu, and K. Q. Weinberger, "Simplifying graph convolutional networks," *arXiv preprint arXiv:1902.07153*, 2019.

[7] B. Settles, "Active learning literature survey," 2009.

[8] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web." Stanford InfoLab, Tech. Rep., 1999.

[9] J.-X. Zhang, D.-B. Chen, Q. Dong, and Z.-D. Zhao, "Identifying a set of influential spreaders in complex networks," *Scientific reports*, vol. 6, p. 27823, 2016.

[10] M. Hopwood, P. Pho, and A. V. Mantzaris, "Exploring the value of nodes with multicommunity membership for classification with graph convolutional neural networks," *Information*, vol. 12, no. 4, p. 170, 2021.

[11] A. K. McCallum, K. Nigam, J. Rennie, and K. Seymore, "Automating the construction of internet portals with machine learning," *Information Retrieval*, vol. 3, no. 2, pp. 127–163, 2000.

[12] C. L. Giles, K. D. Bollacker, and S. Lawrence, "Citeseer: An automatic citation indexing system," in *Proceedings of the third ACM conference on Digital libraries*, 1998, pp. 89–98.

[13] P. Sen, G. Namata, M. Bilgic, L. Getoor, B. Galligher, and T. Eliassi-Rad, "Collective classification in network data," *AI magazine*, vol. 29, no. 3, pp. 93–93, 2008.

[14] J. McAuley, C. Targett, Q. Shi, and A. Van Den Hengel, "Image-based recommendations on styles and substitutes," in *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, 2015, pp. 43–52.

[15] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *science*, vol. 286, no. 5439, pp. 509–512, 1999.

[16] F. Wu, T. Zhang, A. H. d. Souza, C. Fifty, T. Yu, and K. Q. Weinberger, "Simplifying Graph Convolutional Networks," *36th International Conference on Machine Learning, ICML 2019*, vol. 2019-June, pp. 11 884–11 894, 2 2019. [Online]. Available: http://arxiv.org/abs/1902.07153

[17] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[18] P. Pho and A. V. Mantzaris, "Regularized simple graph convolution (sgc) for improved interpretability of large datasets," *Journal of Big Data*, vol. 7, no. 1, pp. 1–17, 2020.

[19] F. Xia, W. Wang, T. M. Bekele, and H. Liu, "Big scholarly data: A survey," *IEEE Transactions on Big Data*, vol. 3, no. 1, pp. 18–35, 2017.

[20] E. Olshannikova, T. Olsson, J. Huhtamäki, and H. Kärkkäinen, "Conceptualizing big social data," *Journal of Big Data*, vol. 4, no. 1, pp. 1–19, 2017.

[21] C. Schmidt, R. Muench, F. Schneider, S. Breitenbach, and A. Carolus, "Generation "always on" turned off. effects of smartphone separation on anxiety mediated by the fear of missing out," in *International Conference on Human-Computer Interaction*. Springer, 2018, pp. 436–443.

[22] N. Ahmed, H. Rahman, and M. I. Hussain, "A comparison of 802.11 ah and 802.15. 4 for iot," *Ict Express*, vol. 2, no. 3, pp. 100–102, 2016.

[23] B. Liu, Z. Yan, and C. W. Chen, "Medium access control for wireless body area networks with qos provisioning and energy efficient design," *IEEE transactions on mobile computing*, vol. 16, no. 2, pp. 422–434, 2016.

[24] T. Tassier, "Labor market implications of weak ties," *Southern Economic Journal*, pp. 704–719, 2006.