2021 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, OCT. 25–28, 2021, GOLD COAST, AUSTRALIA

SCALABLE COMMUNITY DETECTION IN THE DEGREE-CORRECTED STOCHASTIC BLOCK MODEL

Yicong He[⋆]*, Andre Beckus*[⋆] *and George K. Atia*[⋆†]

[⋆]Department of Electrical and Computer Engineering
[†]Department of Computer Science
University of Central Florida, Orlando, FL 32816 USA
yicong.he@ucf.edu, abeckus@knights.ucf.edu, george.atia@ucf.edu

2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP) | 978-1-7281-6338-3/21/$31.00 ©2021 IEEE | DOI: 10.1109/MLSP52302.2021.9596377

ABSTRACT

Community detection aims to partition a connected graph into a small number of clusters. The Degree-Corrected Stochastic Block Model (DCSBM) is one popular generative model that yields graphs with varying degree distributions within the communities. However, large computational complexity and storage requirements of existing approaches for DCSBM limit their scalability to large graphs. In this paper, we advance a scalable framework for DCSBM, in which the full graph is first sub-sampled by selecting a small subset of the nodes, then a clustering of the induced subgraph is obtained, followed by low-complexity retrieval of the global community structure from the clustering of the graph sketch. To sample the underlying graph, we introduce a family of sampling schemes that capture local community structures using metrics derived from the average neighbor degrees, which are shown to achieve the twin objective of sampling from low-density clusters and identifying high-degree nodes within each cluster. The proposed approach can perform on par with full scale clustering while affording substantial complexity and storage gains as demonstrated through experiments using both synthetic and real data.

1. INTRODUCTION

The study of network representations of various phenomena is at the heart of modern network science, with important applications in social science [1, 2], biology [3] and telecommunication [4, 5]. In this realm, community detection has emerged as a useful means for identifying clustering structures intrinsic to networks [6], which could offer insight into how such networks are organized, help uncover some of their important properties, and advance our understanding of the underlying social, natural and physical phenomena.

The stochastic block model (SBM) is a popular generative model of random graphs containing communities in view of its simplicity and tractability. While the SBM is key to numerous noteworthy studies on community detection [7], its value is limited in practice due to the inherent assumption that nodes belonging to the same community are stochastically equivalent. Indeed, many real-world graphs possess a power law degree distribution for which SBM provides a poor fit. This motivated the development of alternative models such as the Degree-Corrected SBM (DCSBM) [8], the Popularity Adjusted Block Model [9], and the LFR benchmark [10], which have better capacity to capture the node degree variability within and between communities typical of real-world data.

There exist several algorithms for community detection in degree-corrected models, including SCORE [11], the Convexified Modularity Maximization (CMM) algorithm [12], the Conditional Pseudo-Likelihood (CPL) algorithm [13] and Weighted K-medians Clustering (WKC) [14]. While many such algorithms perform well under the DCSBM, and some have provable guarantees, they are not viable for clustering large scale networks given their high computation and storage complexities. For example, although the success of CMM in DCSBM is well-documented, its complexity scales superlinearly with the graph size – it is of order $\mathcal{O}(n^6)$ for a graph of size n.

In this paper, we advance a scalable framework for community detection in DCSBM, which extends previous sketch-based randomized frameworks that focused on graphs from the SBM (e.g., [15]). In order to preserve the underlying community structure in a small graph sketch, we develop a novel sampling method termed SAND, which is a family of graph sampling schemes that leverage local clustering properties of the graph, shown to simultaneously sample hubs in the graph and pick enough nodes with high degrees from the low density clusters. The appellation is due to the use of metrics derived from the Average Neighbor Degree (AND) for sampling, which measures the degree of a node relative to the sum of the degrees of its neighbors. We analyze the behavior of the local properties used and show that the resulting sampling

This work was supported in part by NSF CAREER Award CCF-1552497 and NSF Award CCF-2106339.

schemes achieve the desired objectives for successful clustering, through a study of the properties of the sketches they produce in relation to the phase transitions of different clustering algorithms. Further, we propose sequential retrieval and refinement procedures to extrapolate the clustering results to the whole graph and enhance the classification accuracy, respectively. We demonstrate our findings through a set of experiments conducted on both synthetic and real data.

2. DATA MODEL AND SCALABLE FRAMEWORK

2.1. Degree-corrected stochastic block model (DCSBM)

Data Model (DCSBM) [8]: The graph $G = (V, E)$ with vertex and edge sets V and E consists of $|V| = N$ nodes partitioned into r disjoint clusters $C_1, C_2, ..., C_r$. The size of cluster C_i is denoted n_i. Each pair of distinct nodes $i \in C_a$ and $j \in C_b$ are connected with probability $\theta_i \theta_j B_{ab}$, where θ_i is the degree heterogeneity parameter of node i, and $\mathbf{B} \in \mathbb{R}_+^{r \times r}$ is the connectivity matrix of the clusters. The vector of heterogeneity parameters for all nodes is denoted $\boldsymbol{\theta}$. Unless otherwise noted, we assume the common identifiability condition $\max_{i \in C_a} \theta_i = 1, a \in 1, \ldots, r$ to remove any ambiguity regarding the scalings of $\mathbf{B}$ and $\boldsymbol{\theta}$.

2.2. Sketch-based community detection framework

We introduce our scalable sketch-based framework for community detection under the DCSBM, which consists of four steps. First, we obtain a sketch of the graph G by sampling a small subset $S \subset V$, $|S| \ll N$, of the nodes. In the second step, we cluster $G_S = (S, E_S)$, where $G_S = (S, E_S)$ is the induced subgraph of G on S, using a clustering algorithm of our choice. Subsequently, the results of the clustering are extrapolated to each node in the whole graph based on the number of connections the node has to identified clusters in the sketch. In an optional fourth step, a final refinement of the final label assignment is performed on the full graph to further reduce the misclassification rate.

Our framework is related to [15, 16], in that the complex clustering operation is only applied to a small subgraph of the original graph, based on which the global community structure is subsequently obtained. However, these works focused on graphs from the SBM or its heterogeneous variant, which induce clustered graphs with uniform node degree distributions within the communities. Therefore, while their approach is powerful at reducing the computational burden of clustering large graphs when the nodes belonging to a cluster are statistically indistinguishable, it is of little use when the nodes within communities have disparate node degrees as with DCSBM graphs.

3. PROPOSED METHOD FOR DCSBM

3.1. Desirable sketch properties for DCSBM

The analysis in [17] shows that dense networks can yield good asymptotic clustering results. This fact has also been verified through simulations [13, 17]. On the other hand, clustering algorithms typically scale super-linearly in the graph size, thus a small sketch size is desirable to prevent the clustering time from growing unreasonably large. Apart from sampling nodes with large degree in each cluster, some algorithms are sensitive to the presence of imbalanced cluster sizes in the sketch, as analyzed in [18].

We use a simple DCSBM model to verify the previous intuition. The θ_i parameter is drawn independently from the discrete probability distribution $P(\theta_i = 1) = m, P(\theta_i = 0.2) = 1 - m$, i.e., the nodes will have only two possible values for the expected degree. The connectivity matrix $\mathbf{B}$ is

$$\mathbf{B} = \begin{bmatrix} 0.7 & 0.4 & 0.1 & 0.1 \\ 0.4 & 0.7 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.3 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.3 \end{bmatrix} \quad (1)$$

such that the first two clusters are more dense, while the third and fourth clusters are more sparse. The size of the graph is set to $N = 800$ with cluster sizes $n_1 = n_2 = (N - 2n^*)/2$, $n_3 = n_4 = n^*$.

Fig. 1 shows the phase transition of the F-$score$ for four different graph clustering algorithms. It can be seen that SCORE, WKC and CMM require enough nodes from the low-density cluster to obtain good clustering performance, especially when m is small, i.e., the graph density is low.

Further, we illustrate the drawbacks of the current sampling methods, including uniform random sampling (URS), sampling inversely proportional to node degree (SPIN) [15], degree-based random node sampling (DRN), and forest fire sampling (FF) [19, 20]. A graph is generated using the same bi-degree DCSBM model above. The density parameter $m = 0.2$ and the size of the graph $N = 8000$ with equal cluster sizes $n_i = 2000, i = 1, \ldots, 4$. We sample $N' = 800$ nodes, yielding an induced subgraph of the same size as the one in the previous example. The parameter $\hat{m}$ in the sampled sketch is estimated by the fraction of high degree nodes in each cluster. The size $\hat{n}^*$ is estimated by taking the average of sizes of the two low-density clusters in the sketch. The points $(\hat{m}, \hat{n}^*)$ for different sampling methods are marked in Fig. 1. As shown, the current sampling methods do not consistently produce sketches having the aforementioned desirable properties, and therefore often lead to poor clustering results.

3.2. Proposed sampling method

In this section, we develop and characterize a family of flexible sampling schemes, which can achieve the twin objective of sampling enough hubs with high degree and a sufficient number of representatives from the low-density communities.

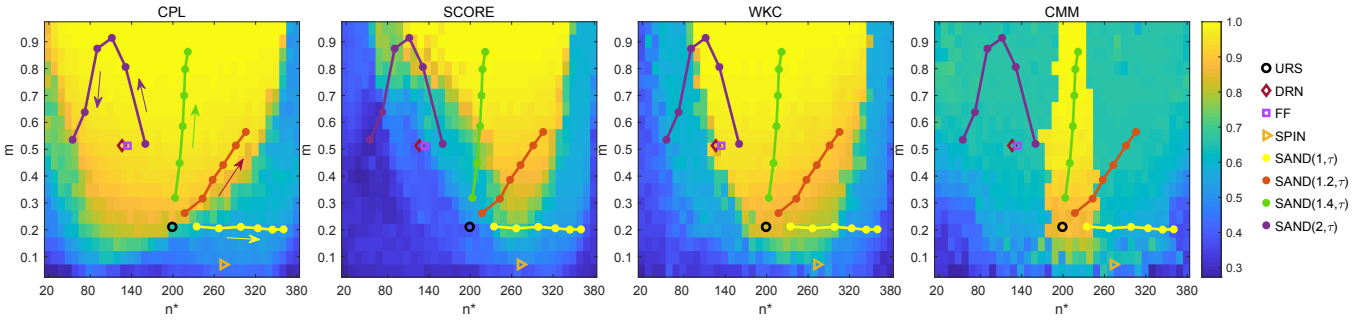

Fig. 1. Phase transitions of different graph clustering algorithms displaying the F-*score* as function of the graph density and the low-density cluster size. Left to right: CPL, SCORE, WKC and CMM. The estimated $(\hat{m}, \hat{n}^*)$ for different sampling methods are marked. For SAND, the values of τ are varied along the direction of the arrow from 1 to 6.

Similar to SBM, in the DCSBM, the degree can be used to estimate *local* information about the degree heterogeneity θ_i. However, it provides little information about the *global* cluster structure. To glean information between clusters, we instead consider the *neighbors* of node i and their degrees. Specifically, we sample the nodes according to the inverse of the average neighbor degree, i.e., with probability proportional to

$$w_i' = \frac{d_i}{\sum_{j\in\mathcal{N}_i} d_j}, \tag{2}$$

where $\mathcal{N}_i$ denotes the neighbors of node i.

For the analysis in this section, we use the condition $\sum_{i\in C_a}\theta_i = 1, a \in 1,\dots,r$ in place of that of Section 2.1. This condition scales $\mathbf{B}$ and $\boldsymbol{\theta}$ such that $\mathbf{B}$ has a convenient interpretation. In particular, B_{ab} becomes the expected number of edges between clusters a and b for $a \neq b$, and B_{aa} is twice the expected number of edges within cluster a. We refer to the diagonal entries of $\mathbf{B}$ using the notation $p_a = B_{aa}$.

Now, observe that the expected value of the numerator of (2) is

$$\mathbf{E}[d_i] = \theta_i H_a - \theta_i^2 B_{aa} \approx \theta_i H_a, i \in C_a, \tag{3}$$

where $H_a := \sum_{b=1}^r B_{ab}$. Likewise, because each edge in the graph is generated independently, the expected value of the denominator of (2) is

$$\mathbf{E}\left[\sum_{j\in\mathcal{N}_i} d_j\right] = \theta_i \sum_{b=1}^r S_b H_b B_{ab}, i \in C_a \tag{4}$$

where $S_a := \sum_{i\in C_a}\theta_i^2$.

To proceed in studying this sampling strategy, we make a simplifying assumption about the parameters of the DCSBM model, namely, we assume that $S_1 = S_2 = \cdots = S_r = s$. This occurs, for example, when the clusters have both equal size and the same distribution of θ_i. We additionally assume that $B_{ab} = q$ for all $1 \le a < b \le r$. With these assumptions, (3) and (4) reduce to

$$\mathbf{E}[d_i] = \theta_i t_a, \;\; \mathbf{E}\left[\sum_{j\in\mathcal{N}_i} d_j\right] = \theta_i s\left(p_a t_a + q\sum_{\substack{b=1\\ b\neq a}}^r t_b\right), \tag{5}$$

where $t_a = \sum_{b=1}^r B_{ab} = p_a + (r-1)q$.

Now, suppose that the graph is "average" in the sense that d_i and $\sum_{j\in\mathcal{N}_i} d_j$ take their expected values. Then, the probability of sampling node i in this graph is proportional to

$$\widehat{w}_i' = \frac{1}{s\left(p_a + q\sum_{\substack{b=1\\ b\neq a}}^r \frac{t_b}{t_a}\right)} \tag{6}$$

Noting that s is constant, then $\hat{w}_i$ is based solely on the number of edges in the clusters. In fact, if q is small, then the probability of sampling a node from cluster a is roughly inversely proportional to the number of edges inside the cluster, i.e., there is a higher probability of sampling from clusters with fewer edges. This will tend to produce a sketch where all clusters have an equal number of edges – exactly the desirable property we seek.

Now, we consider the situation when q is not close to zero, and seek to understand if this strategy still favors clusters with fewer edges. Let $p_{max} = \max\{p_a\}_{a=1}^r, p_{min} = \min\{p_a\}_{a=1}^r$. Then we can show that if

$$\begin{aligned} q < &\frac{(r-2)(2p_{min}-p_{max})}{2(r-1)} \\ &+ \frac{\sqrt{(r-2)^2(2p_{min}-p_{max})^2 + 4(r-1)p_{min}^2}}{2(r-1)}, \end{aligned} \tag{7}$$

then $p_a > p_b$ $(1 \le a, b \le r, b \neq a)$ implies that $\widehat{w}_a' < \widehat{w}_b'$. Therefore, so long as the sufficient condition is met, the sampling will still prefer clusters with fewer edges, thus preserving this desirable feature.

Note that (6) is independent of θ_i. This means that the sampling probability *within* each cluster is equal, i.e., the nodes within each cluster are sampled uniformly at random. As shown earlier, a second goal of our sampling method should be to sample higher degree nodes from within clusters. We now describe how (6) can be modified to accomplish this goal. Specifically, we can square the numerator such that we sample with probability proportional to

$$w_i'' = \frac{d_i^2}{\sum_{j\in\mathcal{N}_i} d_j} \tag{8}$$

Returning to the ideal graph having expected degrees we have

$$\widehat{w}_i'' = \frac{\theta_i t_a}{s\left(p_a + q\sum_{\substack{b=1\\ b\neq a}}^{r} \frac{t_b}{t_a}\right)} \tag{9}$$

In the case that q is very small, we will have $\hat{w}_i'' \approx \theta_i/s$, thus the sampling probability of node i will be directly proportional to θ_i, and completely independent of the number of edges in the cluster. This will in turn lead to more high degree nodes being sampled, as is desired.

Under the given assumptions, (6) and (9) represent two extremes, one of which tends to sample nodes based on the number of edges in their parent cluster, and the other which tends to favor high degree nodes, regardless of the cluster to which they belong. We now propose a parameterized sampling technique which inherits benefits from both (2) and (8). Given parameters x and τ, this technique samples node i with probability proportional to

$$w_i = \left(\frac{d_i^x}{\sum_{j\in\mathcal{N}_i} d_j}\right)^{\tau}. \tag{10}$$

We give this proposed sampling method the name "Sampling based on variants derived from the Average Neighbor Degree" (SAND).

When $\tau = 1$, parameter x allows us to interpolate between the two strategies. If we set $x = 1$ then $w_i = w_i'$, whereas if $x = 2$ then $w_i = w_i''$, while setting $\tau > 1$ with $x = 1$ further pushes the sampling in favor of clusters with fewer edges. On the other hand, setting $\tau > 1$ with $x = 2$ further encourages the sampling of large degree nodes.

We now numerically illustrate the behavior of SAND and demonstrate its versatility, especially as compared to existing methods. The lines in Fig. 1 show the estimated $\hat{m}$ and $\hat{n}^*$ using SAND with different parameters x and τ. Each line shows the results for a fixed x with τ varying in the domain [1,6]. The arrows indicate the direction of increasing τ. We note that by varying SAND along its two degrees of freedom, we can produce sketches that cover a surprisingly wide range of cluster sizes and degree distributions. For each of the four clustering algorithms, this allows us to produce a wide variety of sketches which will yield good performance. Of particular note, SAND(1.4,τ) with $1 \leq \tau \leq 6$ fully lies in the narrow region where CMM exhibits good performance.

3.3. Sequential retrieval method and refinement

The results of the clustering can be extrapolated to each node in the full graph based on the number of connections that the node has to each cluster in the sketch. But, motivated by the intuition that nodes with larger degrees tend to be clustered more accurately, we develop a retrieval method which sequentially assigns nodes from the large degree to small degree, i.e., as the iterations proceed, the new labeled nodes in the graph will be utilized for deciding the label of the remaining unlabeled nodes.

To further reduce the misclassification rate, we also perform a refinement process on the full graph. The refinement process is similar to Algorithm 2 in [14] which determines the new label for the i-th node by counting the number of neighbors of the i-th node belonging to each cluster normalized by its corresponding cluster size, and then assign the label of node i to the cluster that has the maximum normalized counts.

4. EXPERIMENTS

In this section, we evaluate the performance of the proposed scalable framework using both synthetic and real data. Four algorithms are utilized for clustering: CPL, CMM, SCORE and WKC. To sample the graph sketch, we use the four sampling methods URS, DRN, FF and the proposed SAND. The performance metrics used for evaluation are the clustering accuracy and the normalized mutual information (NMI) between the clustering results and the ground truth labels.

4.1. Synthetic data

Here, we investigate the performance of the proposed framework for DCSBM using synthetic data. A graph is generated from Data Model 1 with $r = 4$ clusters of equal size. The connectivity matrix is set to $\rho\mathbf{B}$, where $\mathbf{B}$ is as in (1). The parameter ρ can be adjusted to obtain graphs with different densities. By decreasing ρ, we can reduce the density of the graph, thus making the clustering problem more difficult. The degree heterogeneity parameters θ_i are drawn independently from a power law distribution with density function $f(\theta) = \frac{\alpha\beta^\alpha}{\theta^{\alpha+1}}\mathbf{1}_{\theta\leq\beta}$, followed by a truncation operation $\theta_i = \min(\theta_i, 1)$. Unless specified otherwise, the parameters $\alpha = 1.6$ and $\beta = 0.2$. For SAND, we set the parameter $\tau = 4$.

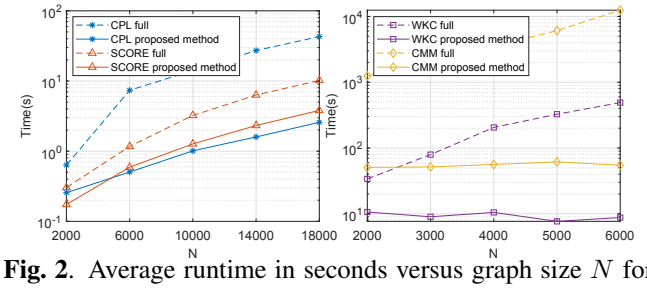

Fig. 2. Average runtime in seconds versus graph size N for proposed method and full graph clustering.

First, we demonstrate the speed improvement afforded by the proposed sketch-based approach compared with clustering of the full graph. SAND is used for sampling, where x and τ are chosen depending on the clustering algorithm. We use $\rho = 1$ and a sketch of size $N' = 1000$. Fig. 2 shows the runtime averaged over 10 Monte Carlo runs versus the graph size N for different clustering algorithms. The proposed framework is shown to considerably improve the running time for

all algorithms, especially WKC and CMM, by using a small sketch for clustering.

Second, we verify the performance of the proposed sampling and retrieval methods. In this experiment, we use $\rho = 0.6$ and graph size $N = 8000$. CMM is used to cluster the graph sketch produced by the different sampling methods. The NMI of the estimated sketch clusters are averaged over 20 Monte Carlo runs and shown in Fig. 3 (left) for different values of the sample size N'. SAND$(1.4, 4)$ yields the best performance, which agrees with our earlier analysis in Section 3.2. Further, we evaluate the performance of the proposed sequential retrieval and refinement procedures. The average NMI of the cluster estimates in the full graph are shown in Fig. 3 (right), again as a function of the sample size N'. We can see that using the proposed sequential retrieval scheme leads to significantly higher NMI than with the retrieval method of [16]. The values are further improved if we include the refinement procedure. Finally, we

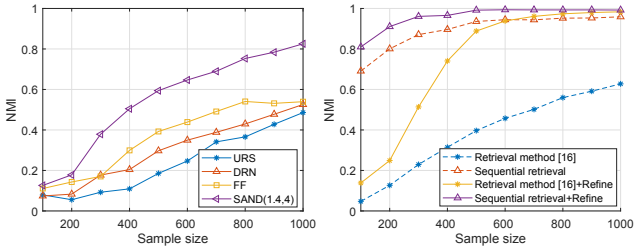

Fig. 3. Left: Average NMI of the sketch for different sampling methods versus sample size. Clustering is performed using CMM. Right: Average NMI of the whole graph with different retrieval methods versus sample size. SAND(1,4,4) is used for sampling and CMM for clustering.

compare the performance of the sketch-based approach with full-scale graph clustering. The graph parameters are set to $N = 8000, \rho = 1, \beta = 0.1$. The sample size N' is set to 1000 for CPL, WKC and CMM, and 2000 for SCORE. For sampling, SAND(2,4) is used for CPL and SCORE, while SAND(1.4,4) is used for WKC and CMM. Fig. 4 shows the NMI for different values of the parameter α of the density function. For CPL, as the network becomes more dense (i.e., when α decreases), the proposed sketch-based framework can improve over directly clustering the full graph. For SCORE, the sketch-based approach outperforms full-scale clustering when the network becomes more sparse, since the sampling method can still produce a relatively dense sketch. For WKC and CMM, the proposed framework performs on par with full-scale clustering.

4.2. Facebook dataset

The Facebook network dataset from [21, 22] consists of data from 100 US universities and a snapshot of all the "friendship" links between the users within each university in September 2005. The dataset also contains several node

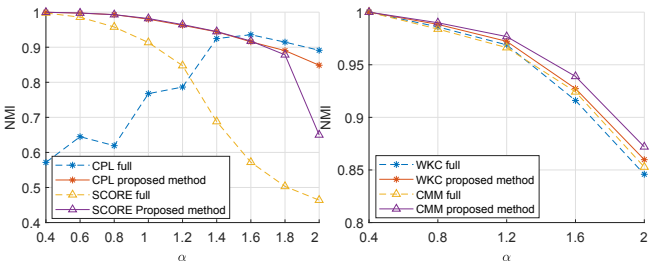

Fig. 4. Average NMI for different clustering algorithms versus parameter α.

attributes, including gender, dorm, graduation year and academic major of each user. In [12], CMM has shown superior performance for Simmons College, which has 1137 nodes and 24,257 undirected edges. In this work, we report on results from the friendship graph of two large networks, 'Brown' and 'Georgetown'. Specifically, we use the year of graduation as the community structure, with the graph induced by the nodes with graduation year between 2006 and 2009. The largest connected component is utilized in the experiment. Finally, we get 5192 and 5676 number of nodes for Brown and Georgetown, respectively. While the number for the two network is 211081 and 155113.

Table 1. Performance comparison on modified networks

Sampling method	Brown		Georgetown	
	Acc(%)	NMI(%)	Acc(%)	NMI(%)
URS	71.25	34.12	72.33	34.76
DRN	66.58	31.17	70.26	31.53
FF	74.09	43.15	65.27	35.32
SAND(1.5,2)	**92.98**	**64.43**	**90.86**	**59.11**
SAND(2,2)	68.11	33.05	74.72	34.90
FULL	56.57	28.18	80.94	43.97

To highlight the advantage of SAND in sampling from graphs with varying degree distributions, we modify the two networks of Brown and Georgetown to increase the density variation between clusters. Specifically, we randomly remove 50% of the links of students who graduated in 2008 and 2009. Community detection is applied to the modified networks using the CMM algorithm with different sampling methods. The average accuracy and NMI are reported in Table 1. The method designated "FULL" refers to direct clustering of the full graph with CMM. One can see that SAND(1.5,2) achieves the best performance with both networks.

5. CONCLUSION

This paper developed a scalable framework for community detection in DCSBM, which captures the degree variability within clusters. We proposed SAND, a parameterized fam-

ily of sampling schemes capable of simultaneously sampling high-degree nodes from the graph and picking enough informative nodes from the low density clusters, which yields favorable sketches for multiple clustering algorithms. Further, a sequential retrieval procedure was proposed to extrapolate the labels in the sketch to all the nodes in the graph and a refinement was utilized to further reduce the misclassification rate. Numerical results for both synthetic and real datasets demonstrate the superior performance of the proposed approach and a significant reduction in time complexity.

6. REFERENCES

[1] Kevin S Xu and Alfred O Hero, "Dynamic stochastic blockmodels for time-evolving social networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 4, pp. 552–562, 2014.

[2] Zizi Papacharissi, *A networked self: Identity, community, and culture on social network sites*, Routledge, 2010.

[3] Olaf Sporns, "Structure and function of complex brain networks," *Dialogues in clinical neuroscience*, vol. 15, no. 3, pp. 247, 2013.

[4] Nan Du, Bin Wu, Xin Pei, Bai Wang, and Liutong Xu, "Community detection in large-scale social networks," in *Proceedings of WebKDD and SNA-KDD workshop on web mining and social network analysis*, 2007, pp. 16–25.

[5] Karsten Steinhaeuser and Nitesh V Chawla, "Community detection in a large real-world social network," in *Social computing, behavioral modeling, and prediction*, pp. 168–175. Springer, 2008.

[6] Santo Fortunato and Darko Hric, "Community detection in networks: A user guide," *Physics reports*, vol. 659, pp. 1–44, 2016.

[7] Emmanuel Abbe, "Community detection and stochastic block models: recent developments," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 6446–6531, 2017.

[8] Brian Karrer and Mark EJ Newman, "Stochastic blockmodels and community structure in networks," *Physical review E*, vol. 83, no. 1, 2011.

[9] Srijan Sengupta and Yuguo Chen, "A block model for node popularity in networks with community structure," *Journal of the Royal Statistical Society: Series B*, vol. 80, no. 2, pp. 365–386, 2018.

[10] Andrea Lancichinetti, Santo Fortunato, and Filippo Radicchi, "Benchmark graphs for testing community detection algorithms," *Phys. Rev. E*, vol. 78, pp. 046110, Oct. 2008.

[11] Jiashun Jin et al., "Fast community detection by score," *The Annals of Statistics*, vol. 43, no. 1, pp. 57–89, 2015.

[12] Yudong Chen, Xiaodong Li, Jiaming Xu, et al., "Convexified modularity maximization for degree-corrected stochastic block models," *The Annals of Statistics*, vol. 46, no. 4, pp. 1573–1602, 2018.

[13] Arash A Amini, Aiyou Chen, Peter J Bickel, Elizaveta Levina, et al., "Pseudo-likelihood methods for community detection in large sparse networks," *The Annals of Statistics*, vol. 41, no. 4, pp. 2097–2122, 2013.

[14] Chao Gao, Zongming Ma, Anderson Y Zhang, Harrison H Zhou, et al., "Community detection in degree-corrected block models," *The Annals of Statistics*, vol. 46, no. 5, pp. 2153–2185, 2018.

[15] Mostafa Rahmani, Andre Beckus, Adel Karimian, and George K Atia, "Scalable and robust community detection with randomized sketching," *IEEE Transactions on Signal Processing*, vol. 68, pp. 962–977, 2020.

[16] A. Beckus and G. K. Atia, "Scalable community detection in the heterogeneous stochastic block model," in *IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP)*, 2019, pp. 1–6.

[17] Yunpeng Zhao, Elizaveta Levina, Ji Zhu, et al., "Consistency of community detection in networks under degree-corrected stochastic block models," *The Annals of Statistics*, vol. 40, no. 4, pp. 2266–2292, 2012.

[18] Shuqin Zhang and Hongyu Zhao, "Community identification in networks with unbalanced structure," *Physical Review E*, vol. 85, no. 6, pp. 066114, 2012.

[19] Jure Leskovec and Christos Faloutsos, "Sampling from large graphs," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2006, pp. 631–636.

[20] Peter Ebbes, Zan Huang, Arvind Rangaswamy, Hari P Thadakamalla, and ORGB Unit, "Sampling large-scale social networks: Insights from simulated networks," in *18th Annual Workshop on Information Technologies and Systems, Paris, France*, 2008.

[21] Amanda L Traud, Eric D Kelsic, Peter J Mucha, and Mason A Porter, "Comparing community structure to characteristics in online collegiate social networks," *SIAM review*, vol. 53, no. 3, pp. 526–543, 2011.

[22] Amanda L Traud, Peter J Mucha, and Mason A Porter, "Social structure of facebook networks," *Physica A: Statistical Mechanics and its Applications*, vol. 391, no. 16, pp. 4165–4180, 2012.