



Improved Community Detection Using Stochastic Block Models

Minhyuk Park, Daniel Wang Feng, Siya Digra, The-Anh Vu-Le,
George Chacko, and Tandy Warnow

Siebel School of Computing and Data Science, University of Illinois
Urbana -Champaign, Urbana, IL 61801, USA
{minhyuk2,chackoge,warnow}@illinois.edu

Abstract. Community detection approaches resolve complex networks into smaller groups (communities) that are expected to be relatively edge-dense and well-connected. The stochastic block model (SBM) is one of several approaches used to uncover community structure in graphs. In this study, we demonstrate that SBM software applied to various real-world and synthetic networks produces poorly-connected to disconnected clusters. We present simple modifications to improve the connectivity of SBM clusters, and show that the modifications improve accuracy using simulated LFR networks.

Keywords: Connectivity · Stochastic Block Model · Clustering

1 Introduction

Community detection is a task in which nodes of a network are partitioned into subsets, each called a community or a cluster (the terms are interchangeable in this manuscript) [2, 12]. Communities do not have to cover an entire network [11], and may also be overlapping [5, 9, 23].

Many community detection methods are based on optimization criteria that reflect one or more of the following properties: preference for clusters that are dense and so have many intra-cluster edges; that are separated from the rest of the network and so have relatively fewer inter-cluster edges; and finally are well-connected, meaning that they do not have small edge cuts (i.e., sets of edges where deleting the edges but not the endpoints separates the cluster into two pieces) [15, 21].

Despite the natural expectation that clusters should be well-connected, this property does not result from some clustering methods [21, 25]. We have previously reported that the Leiden algorithm [21], Infomap [19], Iterative-K-core Clustering (IKC) [22] and Markov Clustering (MCL) [1] community detection algorithms produce clusters that are not well connected [15]. Moreover, [21] documented that the Louvain algorithm can produce disconnected clusters, i.e., clusters that have two or more connected components.

Modifying such clusters to improve connectivity is a logical remediation, and the Connectivity Modifier [15] is one such method that recursively modifies an input clustering to ensure that all final clusters meet a user-provided well-connectedness criterion that depends on the size of the smallest edge cut for the cluster. For the default setting for this criterion, [15] said that a cluster would be considered “well-connected” if the number of edges in the smallest edge cut was strictly greater than $\log_{10}(n)$, where n is the number of nodes in the cluster, and otherwise it was considered poorly connected. While other thresholds can be considered, we use the same setting in this study since it is a very slow-growing function, and thus provides a very mild constraint.

Here, we report on a study examining clustering using Stochastic Block Models (SBMs) [10] on both real-world and synthetic networks. On a collection of more than 100 real-world networks, we find that the SBM clustering methods in graph-tool [17] frequently produce disconnected clusters. We explore three techniques for modifying the clustering to improve the connectivity: simply returning the connected components (CC), repeatedly finding and removing small edge cuts until all clusters meet the default criterion to be considered well-connected (WCC), or applying the recursive Connectivity Modifier method to the clustering. We show that these modifications improve accuracy on LFR networks, with the greatest improvement obtained using WCC.

The rest of this manuscript is as follows. In Sect. 2, we describe the experimental study. We present the results of our experiments on real-world and synthetic LFR networks [8] in Sect. 3. We discuss these results in Sect. 4 and conclude in Sect. 5.

2 Materials and Methods

Due to space constraints, full details are provided in the Supplementary Materials [14].

2.1 Networks

Real-World Networks. We collected a set of 122 real-world networks that range in size from 11 to 13,989,436 nodes. Of these, 120 are from the Netzscheuler network catalogue [18] and we also include the Orkut network (3,072,441 nodes) and the Curated Exosome Network (13,989,436 nodes) [15]. The Netzscheuler network set includes 10 small networks with at most 1000 nodes, 103 medium-sized networks between 1000 and 1,000,000 nodes, and 7 networks with at least 1,000,000 nodes (see Supplementary Materials) [14] for the full list of networks). All real-world networks used in this study were pre-processed to remove self-loops and parallel edges and were treated as unweighted and undirected.

Synthetic Networks. We used synthetic networks that were generated using the LFR [8] software for a previous study [15]. These networks were generated

based on parameters obtained from clusterings computed on five real-world networks using the Leiden algorithm [20] optimizing either the modularity criterion [13] or the Constant Potts Model criterion [21] with five different resolution values (0.0001, 0.001, 0.01, 0.1, 0.5). These LFR networks range in size from 34,546 nodes to 3,774,768 nodes. As reported in [15], a few of these LFR networks had a high incidence of disconnected ground-truth clusters and were not suitable for analysis in this study.

2.2 Stochastic Block Models

We used SBM implemented in graph-tool [17] as a clustering method with the option of three different models: degree-corrected [7], non degree-corrected [4], and planted partition [24]. For each network we clustered using SBMs, we selected the model that had the best fit—i.e., the one with the lowest description length—as our preferred SBM model. We refer to that model as the “selected SBM”, and use it in subsequent post-processing treatments.

2.3 Post-processing Treatments to Improve Connectivity

The Connectivity Modifier (CM) [15] pipeline is designed to modify clusterings in order to ensure that all clusters are well-connected and that no cluster is too small. In prior work [15], we found that the CM pipeline typically improved Leiden clustering accuracy on synthetic networks, and that when it reduced accuracy this was due to removing small clusters. Hence, in this study, we have eliminated the filtering of small clusters, and restricted the CM pipeline to modifying clusterings in order to ensure edge-connectivity.

We evaluate the use of this simplified CM approach as well as two other post-processing treatments, each of which takes as input a clustering \mathcal{C} of a network N , and modifies it, if necessary, to ensure some standard for edge-connectivity. The three treatments we study are:

- **CC (Connected Components)**: If a cluster is disconnected, we return each of its connected components as a cluster.
- **WCC (Well-Connected Clusters)**: We modify clusters by repeatedly removing small edge cuts of size at most $\log_{10}(n)$ until each cluster is well-connected. To find small edge cuts, we use VieCut [3]. Each of these pieces is then examined for well-connectedness and further processed, if needed.
- **CM (Connectivity Modifier)**: We apply the inner loop of the Connectivity Modifier pipeline [15]. If a cluster C has a small edge cut, then removal of the edge cut divides C into two subsets, and each of these is then “re-clustered” using the same clustering method used to produce the input clustering \mathcal{C} . These clusters are then added back into the iterative algorithm, which checks each cluster for being well-connected. Each iteration finds and removes small edges cuts, and then reclusters the two sets. The iteration stops when the cluster satisfies the edge-connectivity criterion.

2.4 Evaluation

We report cluster statistics, including percent of clusters that are connected, percent well-connected, and percent poorly connected. We also report cluster size distributions and node coverage after restriction to clusters of size at least two. For synthetic networks, we report accuracy, measured using three standard criteria: Normalized Mutual Info (NMI), Adjusted Rand Index (ARI), and Adjusted Mutual Info (AMI). For all three accuracy criteria, we used the implementation provided by the Scikit-learn library [16].

3 Performance Study and Results

Here we describe the experiments we performed and the results we obtained.

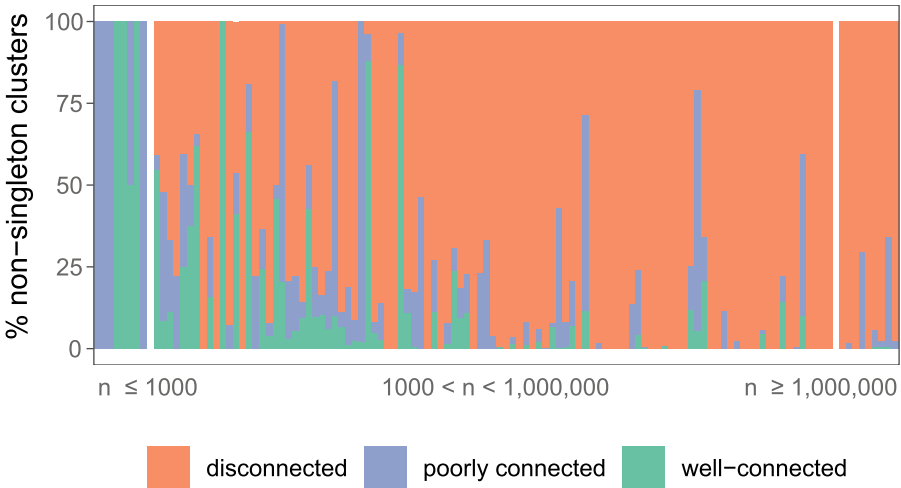


Fig. 1. Experiment 1: Cluster Connectivity of SBM on 120 Real-World Networks. Percentage of disconnected, poorly connected, and well-connected clusters are shown for the selected SBM clustering of 120 real world networks. Each colored bar represents a single network, white bars separate the network groups into small, medium, and large. This figure does not include results for two networks from the initial set of 122 networks, as the selected SBM model returned no non-singleton clusters. The networks are sorted by the number n of nodes

3.1 Experiments

We conducted three experiments:

- Experiment 1: We evaluate the connectivity profile of SBM clusterings on real-world networks.

- Experiment 2: We evaluate the impact of returning the connected components of the clusters of SBM clusterings on real-world networks.
- Experiment 3: We evaluate the impact of our three treatments on clustering accuracy on synthetic networks.

All experiments were performed on the Illinois Campus Cluster with maximum computational resources limit set to 72 h of runtime, 256 GB of RAM, and 16 cores of parallelism.

3.2 Experiment 1: Connectivity of SBMs

In this experiment, we examined the connectivity profile of clusters generated by SBM. Figure 1 shows the cluster connectivity status for the selected SBM model (Sect. 2.2) on each of the networks, which are sorted from left to right by the number of nodes, which range in size from 11 to 13,989,436 nodes. Here, red indicates that the cluster is disconnected, blue indicates poorly connected, and green indicates well-connected. For networks with at most 1000 nodes, clusters are connected, and often well-connected. Above 1000 nodes, however, the clusters in the selected SBM are very often disconnected, and most clusters are disconnected for most of the networks in the upper half of the size range.

3.3 Experiment 2: Impact of Treatments on Real-World Networks

Table 1. Impact of Treatment on Node Coverage on Real-World Networks

For small, medium, and large network groups, the node coverage (i.e., percentage of nodes in non-singleton clusters) is shown for the selected SBM before and after treatment. On one of the medium networks, WCC ran into a memory error with 256GB of RAM, hence the results for that network are omitted from this calculation

clustering	Node Coverage		
	small	medium	large
Selected SBM	62%	100%	100%
Selected SBM - CC	62%	48%	45%
Selected SBM - WCC	55%	36%	25%
Selected SBM - CM	55%	25%	17%

The three post-clustering treatments we apply operate by breaking an input clustering into sub-clusters, and so will change the cluster size distribution, the number of clusters, and even node coverage (i.e., the percentage of nodes in non-singleton clusters). Specifically, if in the process singleton clusters are created, then the node coverage, which is calculated based on non-singleton clusters, can also reduce.

We first examine node coverage (Table 1). Reductions in node coverage are relatively small on the small networks, but all three treatments produce large reductions on the medium and large networks. The largest reductions are for CM, and the smallest are for CC, with WCC in between. However, even CC produces a large reduction in node coverage. Since node coverage is the percentage of nodes in non-singleton clusters, this reduction can only occur because enough nodes are placed in clusters where they do not have any neighbors.

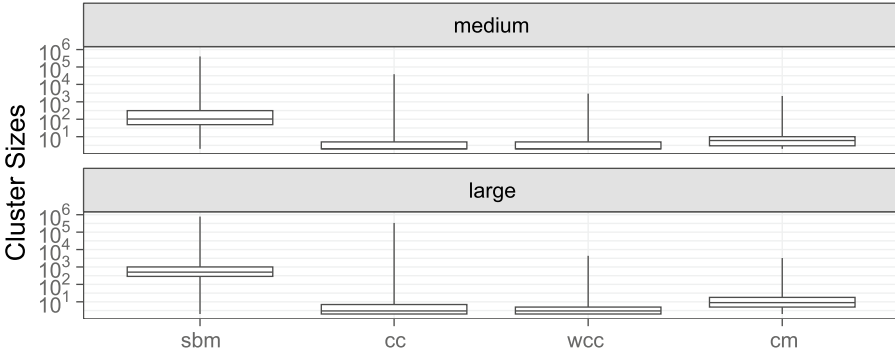


Fig. 2. Experiment 2: Impact of Treatment on Cluster Sizes of Medium and Large Real-World Networks. The distribution of non-singleton cluster sizes is shown as a boxplot for the selected SBM and its treatments. The y-axis is plotted on a log scale. The whiskers indicate the minimum and maximum cluster sizes in all of the networks in the group. Both groups and treatments have minimum cluster size of 2 for SBM clusterings whether treated or not, but differ in the medians and maxes, as follows. Medium group median/max: SBM: 103/403801, SBM+CC: 2/38539, SBM+WCC: 2/2966, SBM+CM: 6/2169. Large group median/max: SBM: 507/777770, SBM+CC: 3/337018, SBM+WCC: 3/4387, SBM+CM: 9/3258

We next examine the impact on cluster size distribution (Fig 2). For both medium-sized networks (top) and large networks (bottom), the median cluster size before treatment is much larger than the final median cluster size after treatment, and this holds for all three treatments. Moreover, the majority of clusters are dramatically reduced in size by the treatments. Even CC, which only modifies the clusters to return connected components, produces a large impact on the cluster size distribution. The cluster sizes seem to be impacted slightly less when CM treatment is applied compared to CC or WCC, both of which have similar impacts on the cluster sizes regardless of network size.

Finally, we examine the impact on the number of non-singleton clusters (Supplementary Materials). All three treatments increase the number of such clusters, and on average CM produced the smallest number of non-singleton clusters, CC produced the next smallest, and then WCC, which produced the most non-singleton clusters.

3.4 Experiment 3: Impact of Treatment on Synthetic Networks

In order to assess the impact of these treatments beyond empirical properties of clusterings, we use synthetic LFR networks with ground truth clusterings to capture the effect treatment has on clustering accuracy. We evaluated NMI, ARI, and AMI accuracies on the LFR networks tested. Recall that some LFR networks had disconnected clusters (i.e., the LFR networks based on CEN clustered using Leiden-CPM with $r = 0.1$ or $r = 0.5$, and the wiki_topcats clustered using Leiden-CPM with $r = 0.5$). On the LFR network for cit_patents with $r = 0.5$, WCC treatment could not produce a clustering within our runtime limit of 72 h when starting with the selected SBM clustering, and so had a “time-out”.

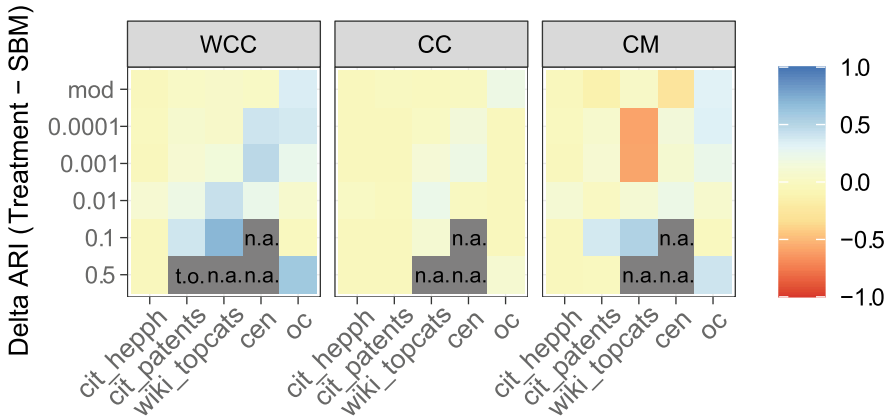


Fig. 3. Experiment 3: Impact of Treatment on ARI Scores of Selected SBM (heatmap). Each LFR network is based on a Leiden clustering of a real-world network, with the column indicating the real-world network and the row specifying the optimization problem (either modularity or CPM for a given resolution value). Blue indicates that post-processing using the corresponding treatment improves ARI accuracy for the clustering method, orange and red indicate that treatment hurts ARI accuracy, and yellow indicates neutral impact. We use “n.a.” to indicate that a network was either not used because of too many disconnected ground-truth clusters or that the LFR software failed to generate the network, and “t.o.” to indicate that WCC failed to complete within 72 h

Fig 3 shows that WCC and CC treatments range from neutral (yellow) to beneficial (blue) with respect to the ARI accuracy of SBM clusterings, but WCC improvements are both more frequent and larger than CC improvements. In comparison, CM can even be detrimental. The relative benefit of WCC over CC and CM holds as well for NMI and AMI (Supplementary Materials), but for those criteria the impact is generally lessened. Overall, therefore, WCC is the preferred treatment for SBM on these networks.

We explore the impact of WCC in greater detail, noting the ARI accuracy for SBM and the final accuracy for SBM-WCC (Fig 4). Note that results are

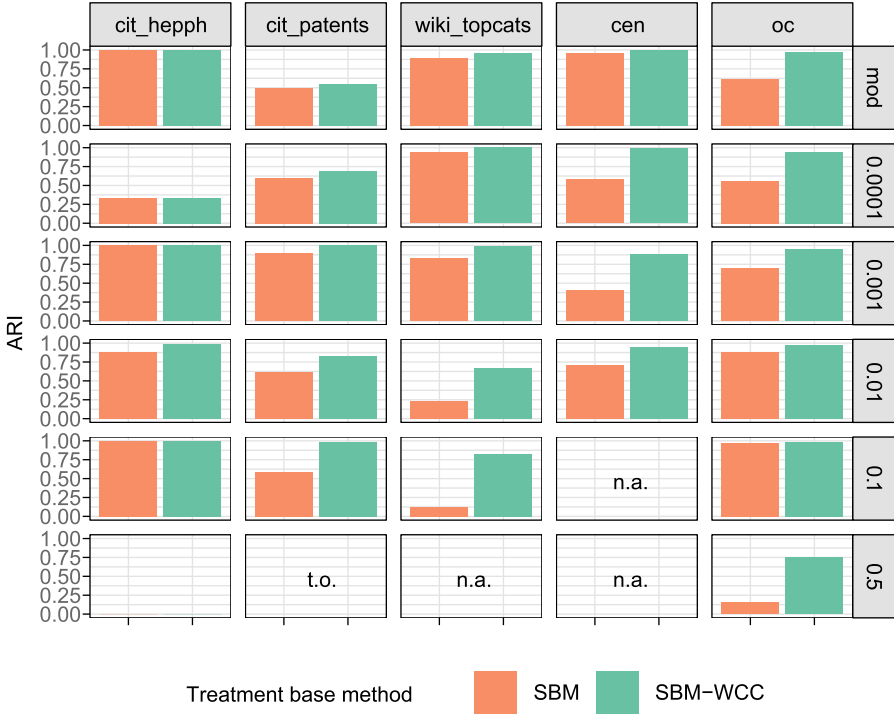


Fig. 4. Experiment 3: Impact of WCC Treatment on ARI Scores of Selected SBM (bar chart). Boxes marked as “n.a.” are for conditions where the LFR networks had too many disconnected ground truth clusters or failed to generate; “t.o.” indicates a failure to complete within 72 h. On the LFR network for cit_hepph with $r = 0.5$, both the selected SBM model and its follow-up WCC yielded 0.0 ARI accuracy

not shown (marked as “n.a.”) for some LFR networks (i.e., the 0.5 CEN, 0.1 CEN, and 0.5 wiki_topcats), because they have many disconnected ground truth clusters or failed to generate, as reported in [15]. On the 0.5 cit_patents network, WCC treatment could not produce a clustering within our runtime limit of 72 h when starting with the selected SBM clustering, and so is marked as a time-out (“t.o.”).

In every case, SBM-WCC is at least as accurate as SBM. Moreover, there are many cases where the benefit from WCC treatment is very large (e.g., Open Citations with resolution value 0.5, the Curated Exosome Network with resolution value 0.001, wiki_topcats with resolution values 0.01 and 0.1). Finally, most of the cases where WCC has at most a small positive impact are for the cases where SBM is already highly accurate, with close to 1.00 ARI accuracy, so that there is little room for improvement.

We also examined the models selected by SBM on the LFR networks. We found that on the LFR networks, the non degree corrected SBM never yielded the lowest description length, and that typically the planted partition model had

the lowest description length (Supplementary Materials). We also saw that the lowest description length model resulted in the best NMI/ARI accuracies on the LFR networks tested.

4 Discussion

Summary of Trends. As seen in Figs. 3 and 4, both CC and WCC post-processing treatments improve accuracy for SBM clusterings on simulated datasets, and WCC is particularly beneficial for accuracy. CM sometimes improves accuracy but sometimes reduces accuracy, and so is not recommended. However, the impact on cluster size and node coverage produced by even the simple CC technique is noteworthy. Our prior work [15] showed that CM improved Leiden clustering accuracy, which is different from what we observe here for SBM clusterings. While the reason for this is not clear, the tendency of SBM to produce disconnected clusters may be part of the explanation.

Impact of the Description Length Formula. Given that the construction of SBMs within graph-tool produces disconnected clusters, we now consider how the code operates. Recall that this approach seeks to generate SBMs that optimize the description length, and this is a minimization problem. Let

- A be the adjacency matrix,
- b be the cluster (block) assignment,
- k be the degree vector (induced by A),
- and e be the edge count matrix (induced by A and b).

In Eq. (1) we provide the formula for the description length of a clustering b of a network given by its adjacency matrix A (i.e., $DL(A, b)$) under the Degree Corrected (DC) model:

$$DL(A, b) = -\log p(A|b, e, k) - \log p(k|b, e) - \log p(b) - \log p(e) \quad (1)$$

Note that the description length is calculated as the sum of various components: the negative logarithm of the model likelihood (i.e., $-\log p(A|b, e, k)$) and the negative logarithm of each of the priors.

In our analyses (Supplementary Materials), we observed that the model likelihood without priors favors connected clusters returned by the CC treatment. In contrast, certain priors heavily penalize having many clusters, leading to a worse description length for the clustering returned by the CC treatment.

We provide an example of this phenomenon on a real-world network, `linux`, in Table 2; in the Supplementary Materials, we show that the trends observed on this network are also observed in the other real-world networks. One clustering is from SBM(DC) (i.e., the degree corrected SBM output) and the second clustering is from SBM(DC)-CC (i.e., the result of running the CC treatment on the degree corrected SBM output). The SBM(DC) clustering has a lower description length than SBM(DC)-CC, and hence the untreated SBM clustering

Table 2. Breakdown of Description Lengths on the linux real-world network

The last row is the sum of the values in the first four rows. The ratio is SBM(DC)-CC SBM(DC), so that values less than 1.0 favor SBM treated by CC and values greater than 1.0 favor untreated SBM. Bold text indicates the preferred clustering.

Quantity	SBM(DC)	SBM(DC)-CC	Ratio
$-\log p(A b, e, k)$	699228.26	315644.88	0.45
$-\log p(k b, e)$	95737.43	45066.47	0.47
$-\log p(b)$	147018.92	256817.11	1.75
$-\log p(e)$	50786.40	1584554.98	31.20
DL(A, b)	992771.01	2202083.44	2.22

is the preferable clustering with respect to the minimization of the description length under the degree corrected model. However, although $-\log p(A|b, e, k)$ and $-\log p(k|b, e)$ for the SBM(DC)-CC clustering are lower, $-\log p(b)$ and $-\log p(e)$ for SBM(DC)-CC clustering are higher, and by a larger magnitude, and hence offset the first two quantities. Moreover, between these two priors, the $-\log p(e)$ component has the bigger impact on this outcome. Furthermore, if this component had *not been included* in the formula, then SBM(DC)-CC would have a lower description length, and would have been favored.

We examined the other 102 networks that had selected DC as the model. For all of these, the $-\log p(e)$ component strongly favored the untreated SBM over the treated SBM. We then examined whether removing the $-\log p(e)$ component of the summation of the description length for both treated and untreated SBM models, to see which model would have been returned. We found that for 80 of the 103 networks in total, removing this component of the summation would have resulted in the treated SBM model having a lower description length than the untreated model. Thus, this specific component of the summation accounts for 77.7% of the cases where the untreated SBM is favored over the treated SBM.

We examine the formula for the negative logarithm of the prior for the edge count matrix, which is given by:

$$-\log p(e) = \log \left(\frac{B(B+1)/2 + E - 1}{E} \right) \quad (2)$$

where B and E are the number of blocks and edges, respectively. As B increases, this value will increase, and does so quickly. This explains why clusterings with a larger number of clusters (such as are produced by running CC, WCC, and CM) have larger description lengths, and hence are less favored.

5 Conclusion

Our study demonstrates that clustering using SBMs is prone to producing disconnected clusters, with the frequency of disconnected clusters increasing as the

network size grows. We show that two simple techniques—CC, which returns the connected components of the clusters, and WCC, which repeatedly removes small edge cuts (based on a mild criterion for “small” that depends on the size of the cluster)—can be used to modify an SBM clustering and tends to improve clustering accuracy on synthetic networks. Moreover, WCC has the strongest improvements in our simulation study. Interestingly, using the Connectivity Modifier [15] under the same mild criterion had variable impact, sometimes improving and sometimes reducing accuracy. Thus, our study provides two simple ways to modify clustering using SBMs that lead to improvements in accuracy, without requiring substantial computational resources.

This study focused on improving clustering quality for SBMs, but did so through essentially *ad hoc* techniques. Future work should investigate how to achieve these improvements and guarantees of connectivity within the model-based framework of SBMs. Other future work includes evaluation using other synthetic networks, such as ABCD [6], and exploring whether other post-processing approaches can lead to larger improvements in accuracy while maintaining scalability.

Funding Information. This work was supported in part by the Illinois-Inspire partnership.

Software. See <https://github.com/MinhyukPark/constrained-clustering> for the CC and WCC codes.

References

1. Dongen, S.V.: Graph clustering via a discrete uncoupling process. *SIAM J. Matrix Anal. Appl.* **30**(1), 121–141 (2008)
2. Fortunato, S., Newman, M.E.J.: 20 years of network community detection. *Nat. Phys.* **18**(8), 848–850 (2022)
3. Henzinger, M., Noe, A., Schulz, C., Strash, D.: Practical minimum cut algorithms. *ACM J. Exp. Algorithmics* **23**, 1–22 (2018)
4. Holland, P.W., Laskey, K.B., Leinhardt, S.: Stochastic blockmodels: first steps. *Soc. Netw.* **5**(2), 109–137 (1983)
5. Jakatdar, A., Liu, B., Warnow, T., Chacko, G.: AOC: assembling overlapping communities. *Quant. Sci. Stud.* **3**(4), 1079–1096 (2022)
6. Kamiński, B., Prałat, P., Thériberge, F.: Artificial benchmark for community detection (ABCD)—Fast random graph model with community structure. *Netw. Sci.* **9**(2), 153–178 (2021). <https://doi.org/10.1017/nws.2020.45>
7. Karrer, B., Newman, M.E.: Stochastic blockmodels and community structure in networks. *Phys. Rev. E-Statistical Nonlinear Soft Matter Phys.* **83**(1), 016107 (2011)
8. Lancichinetti, A., Fortunato, S., Radicchi, F.: Benchmark graphs for testing community detection algorithms. *Phys. Rev. E* **78**(4), 046110 (2008)
9. Lancichinetti, A., Radicchi, F., Ramasco, J.J., Fortunato, S.: Finding statistically significant communities in networks. *PloS One* **6**(4), e18961 (2011)

10. Lee, C., Wilkinson, D.J.: A review of stochastic block models and extensions for graph clustering. *Appl. Netw. Sci.* **4**(1) (2019). <https://doi.org/10.1007/s41109-019-0232-2>
11. Miasnikof, P., Shestopaloff, A.Y., Raigorodskii, A.: Statistical power, accuracy, reproducibility and robustness of a graph clusterability test. *Int. J. Data Sci. Anal.* **15**(4), 379–390 (2023)
12. Newman, M.E.J.: Detecting community structure in networks. *Eur. Phys. J. B* **38**(2), 321–330 (2004). <https://doi.org/10.1140/epjb/e2004-00124-y>
13. Newman, M.E., Girvan, M.: Finding and evaluating community structure in networks. *Phys. Rev. E* **69**(2), 026113 (2004)
14. Park, M., Feng, D.W., Digra, S., Vu-Le, T.A., Chacko, G., Warnow, T.: Supplementary materials for improved community detection using stochastic block models (2024). <https://doi.org/10.5281/zenodo.1334515>
15. Park, M., et al.: Well-connectedness and community detection. *PLOS Complex Syst.* **1**(3), e0000009 (2024)
16. Pedregosa, F., et al.: Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011)
17. Peixoto, T.P.: The graph-tool python library. figshare (2014). <https://doi.org/10.6084/m9.figshare.1164194>, <http://figshare.com/articles/graphtool/1164194>
18. Peixoto, T.P.: The netzscheleuder network catalogue and repository. *Zenodo* **10** **5281** (2020). <https://doi.org/10.5281/zenodo.7839980>. <https://zenodo.org/records/7839981>
19. Rosvall, M., Axelsson, D., Bergstrom, C.T.: The map equation. *Eur. Phys. J. Special Top.* **178**(1), 13–23 (2009)
20. Traag, V.: Leiden algorithm: leidenalg. <https://github.com/vtraag/leidenalg> (2019)
21. Traag, V.A., Waltman, L., Van Eck, N.J.: From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* **9**(1), 1–12 (2019)
22. Wedell, E., Park, M., Korobskiy, D., Warnow, T., Chacko, G.: Center-periphery structure in research communities. *Quant. Sci. Stud.* **3**(1), 289–314 (2022)
23. Xie, J., Kelley, S., Szymanski, B.K.: Overlapping community detection in networks: the state-of-the-art and comparative study. *ACM Comput. Surv.* **45**(4), 1–35 (2013)
24. Zhang, L., Peixoto, T.P.: Statistical inference of assortative community structures. *Phys. Rev. Res.* **2**(4), 043271 (2020)
25. Zhu, Z.A., Lattanzi, S., Mirrokni, V.: A local algorithm for finding well-connected clusters. In: *International Conference on Machine Learning*, pp. 396–404. PMLR (2013)