The Problem of Cluster Connectivity in Community Detection Methods

Tandy Warnow

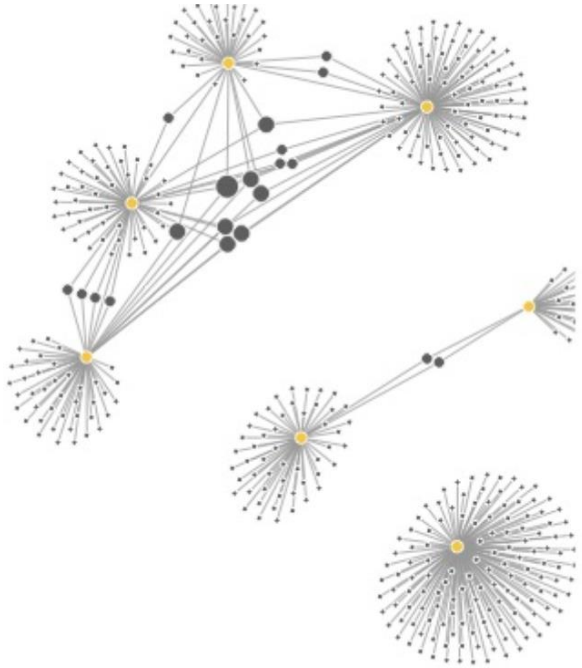
Univ Illinois Urbana-Champaign

Supported by the Insper-Illinois Partnership, Oracle, Digital Science, Google, and the Grainger Foundation

The Scientometrics and Network Science Project, Chacko-Warnow Collaboration

Goals:

1. Understanding the organization of scientific communities, and especially emerging trends in biomedical research
2. Developing novel community detection and community search methods that enable discovery in large networks
3. Developing new methods for understanding community structure in large networks (millions of nodes), including the detection of overlapping communities and evolution of communities over time.



https://tandy.cs.illinois.edu/bibliometrics.html

This talk

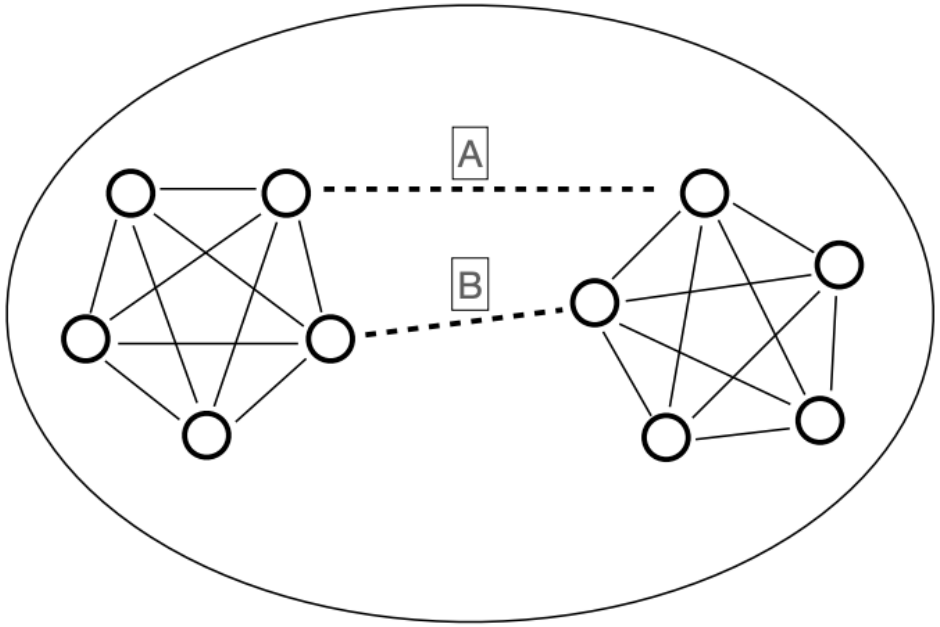
I. Cluster connectivity: what this means, what is known
II. The Connectivity Modifier (Park et al. Complex Networks and their Applications, 2023, and PLOS Complex Systems, in press) – results on Leiden, IKC, MCL, Infomap
III. Clustering using Stochastic Block Models (Park et al., Complex Networks and their Applications 2024)
IV. Synthetic networks using Stochastic Block Models (Anne et al., Complex Networks and their Applications 2024)
V. Conclusions

Overall summary

- Cluster connectivity is important for clustering and synthetic network generation
- Simple ad hoc techniques are helpful
- Rigorous mathematical approaches and models are needed

I. Well-connected = no small edge cut

- Edge cut: set of edges whose removal splits the graph into separate components
- For the graph shown:
  - No single edge removal disconnec the graph
  - An edge cut of size 2: {A,B}
  - Min edge cut size is 2.



Related to “set conductance” of each cluster, several papers in the CS literature (e.g., Kannan et al., JACM 2004; Koutis and Miller SPAA 2008; Zhu et al., ICML 2013)

scientific reports

Explore content | About the journal | Publish with us

nature > scientific reports > articles > article

Article | Open access | Published: 26 March 2019

From Louvain to Leiden: guaranteeing well-connected communities

V. A. Traag, L. Waltman & N. J. van Eck

Scientific Reports **9**, Article number: 5233 (2019) | Cite this article

120k Accesses | **1317** Citations | **222** Altmetric | Metrics

(1) Introduced Leiden algorithm

(2) Demonstrates Louvain produces disconnected clusters

(3) Proves CPM-optimal clusters "well-connected" (based on their definition)

Traag 2019: CPM-optimal clusterings are well-connected

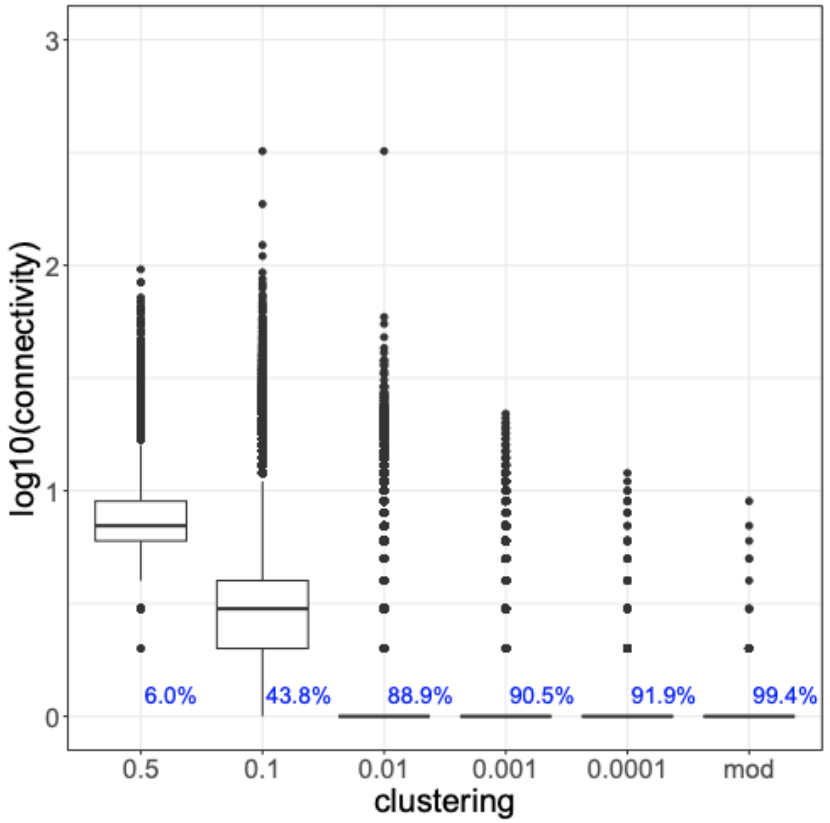
The Constant Potts Model (CPM) optimization score depends on the resolution parameter γ

$$\mathcal{H} = \sum_c \left[e_c - \gamma \binom{n_c}{2} \right]$$

Theorem (rephrased from Traag et al. 2019):
Let C be a cluster in an optimal CPM clustering for resolution parameter γ.
Suppose removing edge set E' splits C into sets X and Y.
Then E' has at least γ |X||Y| edges.

This lower bound depends on γ and is not very meaningful when γ is small

Leiden clusters have small edge cuts



Leiden optimizing either Modularity (mod) or the Constant Potts Model (CPM) for different resolution values.

Blue text in left figure indicates node coverage

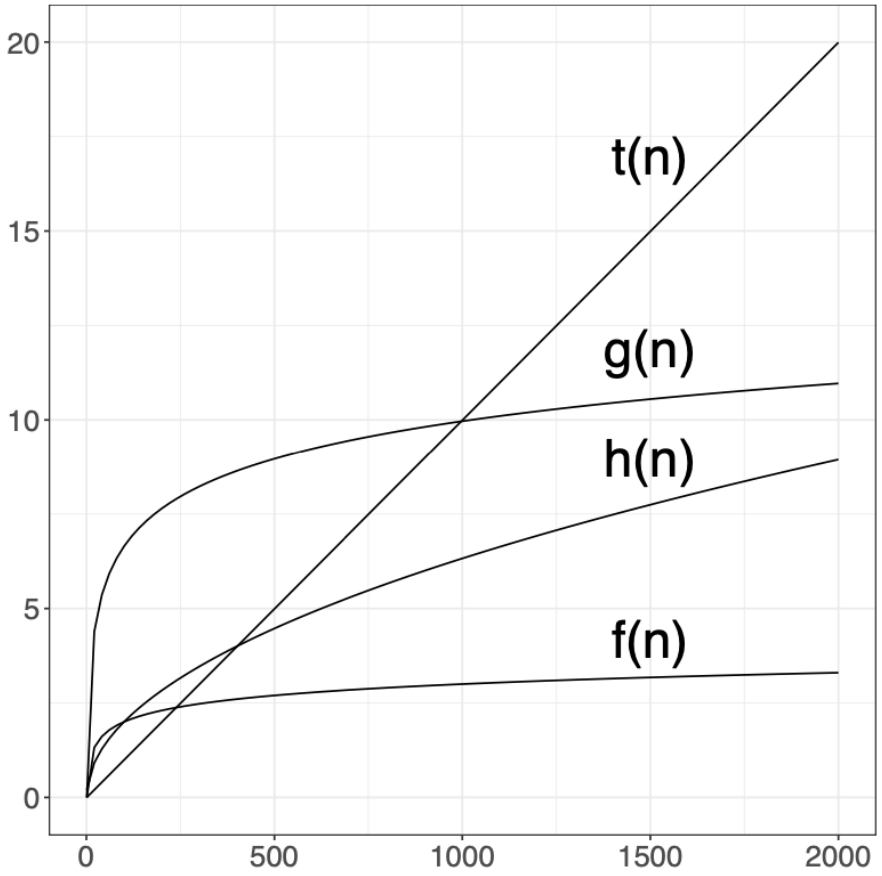
Trade-off between node coverage and edge-connectivity

Figure 1: *Node coverage, connectivity, and size distribution of clusters generated by Leiden optimizing either CPM or modularity on the Open Citations network (75,025,194 nodes).*

II. The Connectivity Modifier

- Park et al. (2023, 2024): Well-Connectedness and Community Detection

Lower bounds for “well-connected” clusters with n nodes



n = cluster size

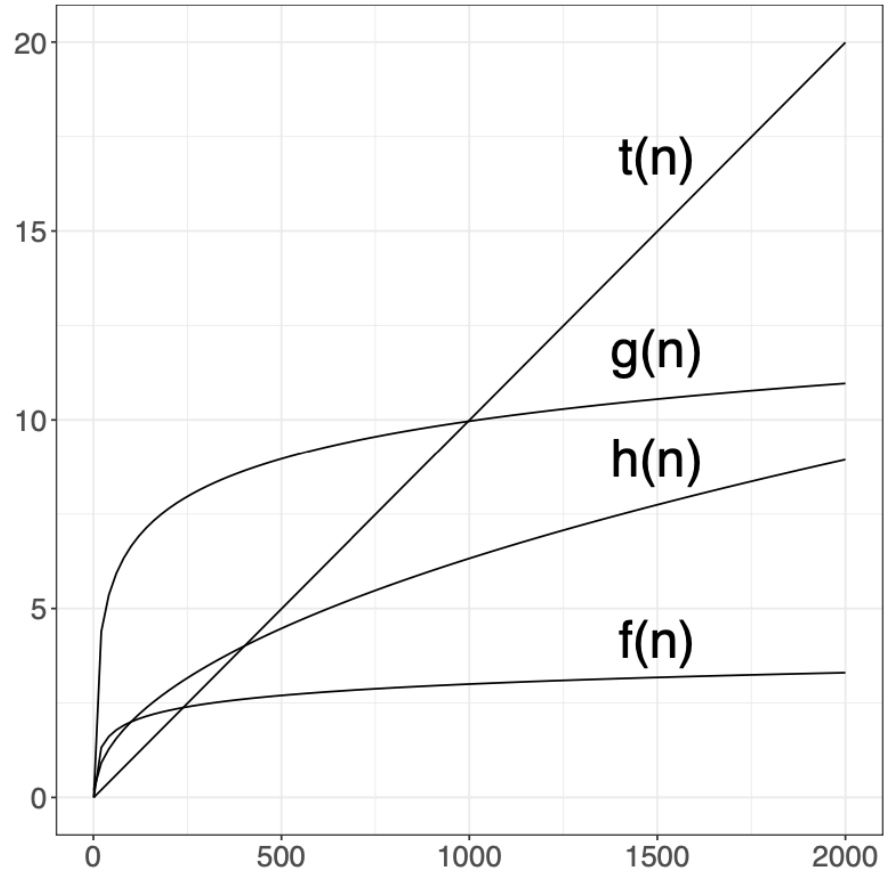
$f(n) = \log_{10} n$

$g(n) = \log_2 n$

$h(n) = (n^{0.5})/5$

t(n) = 0.01(n-1): the guarantee for CPM-optimal clusterings when γ = 0.01

We use f(n) for “well-connected”



n = cluster size

$f(n) = \log_{10} n$

A cluster must have no edge cut of size at most f(n) to be “well-connected”

Our study: networks and community detection methods

network	nodes	edges	avg_deg	ref
Open Citations	75,025,194	1,363,605,603	36.35	(*17*)
CEN	13,989,436	92,051,051	13.16	(*35*)
cit_hepph	34,546	420,877	24.37	(*36*)
cit_patents	3,774,768	16,518,947	8.75	(*36*)
orkut	3,072,441	117,185,083	76.28	(*37*)
wiki_talk	2,394,385	4,659,565	3.89	(*38*)
wiki_topcats	1,791,489	25,444,207	28.41	(*39*)

We also examined synthetic networks based on these networks.

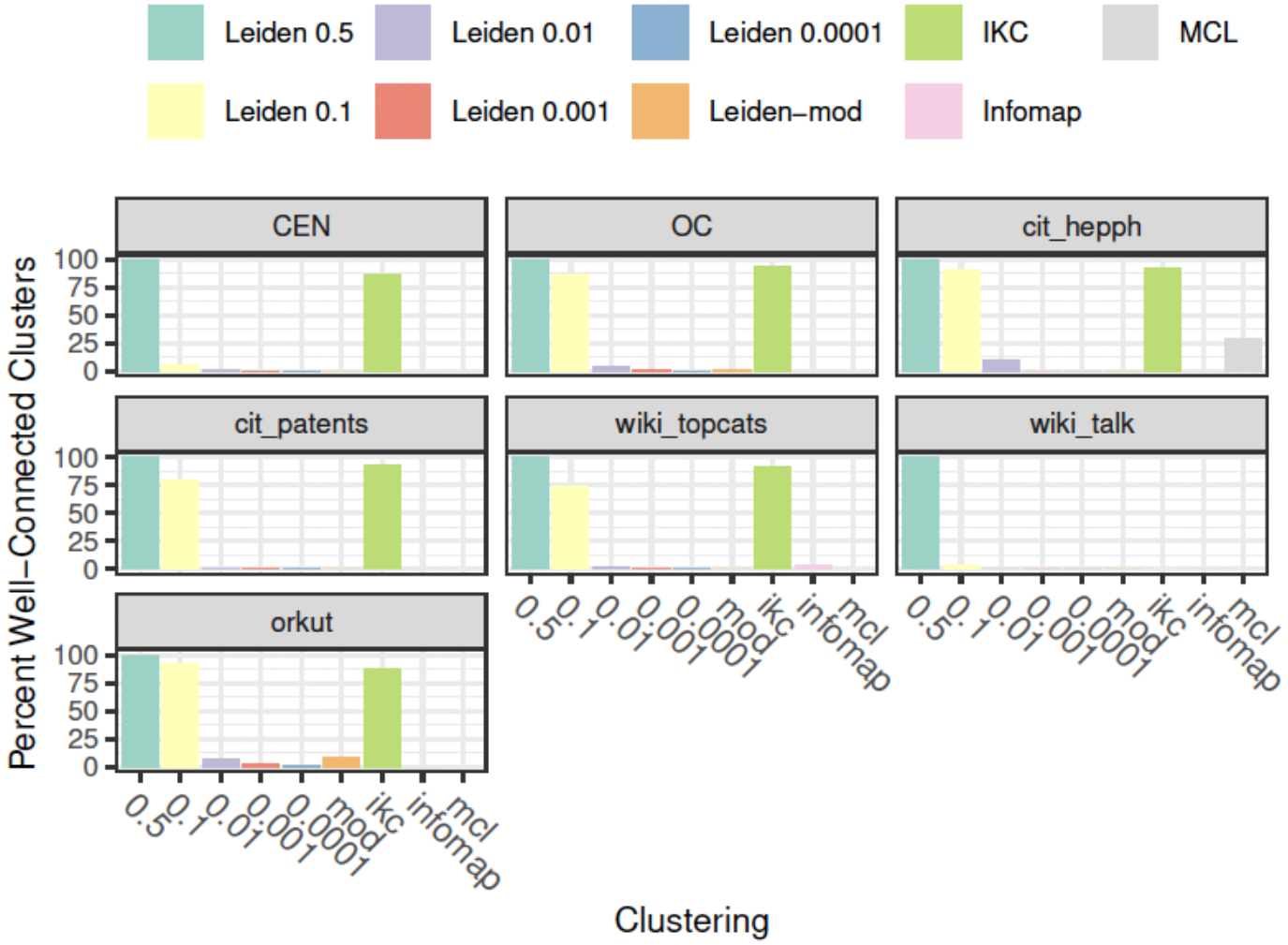
Only Leiden and IKC completed on Open Citations.

IKC had very low node coverage

Community Detection Methods:

- Leiden optimizing Modularity and the Constant Potts Model (CPM)
- Iterative k-core (IKC)
- Markov Clustering (MCL)
- Infomap

Well-connectedness in 7 real-world networks



The Connectivity Modifier (CM) Pipeline

CM reclusters in each iteration, using a selected clustering method

Parameter Defaults:

- Well-connected means min cuts above $\log_{10} n$
- Cluster min size 11

Now includes other clustering methods

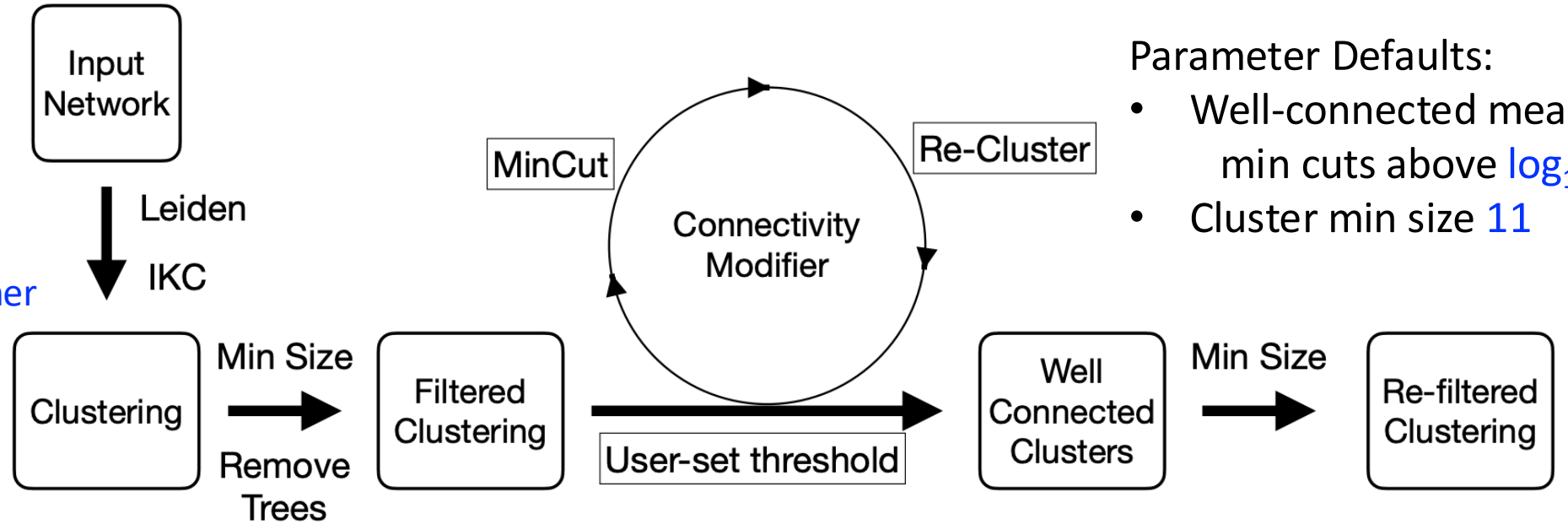
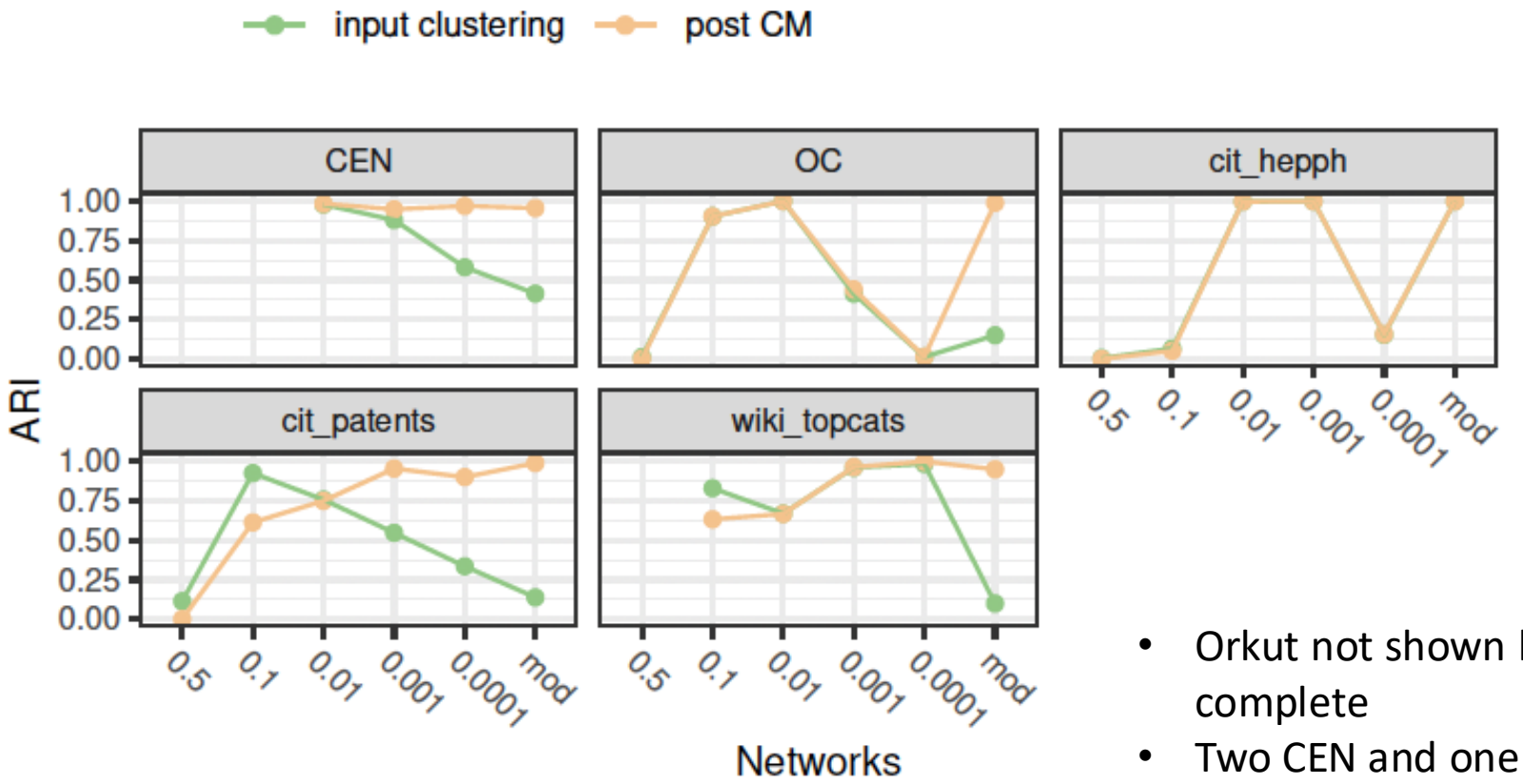


Figure 3: *Connectivity Modifier Pipeline Schematic* The four-stage pipeline depends on user-

CM improves accuracy on synthetic networks

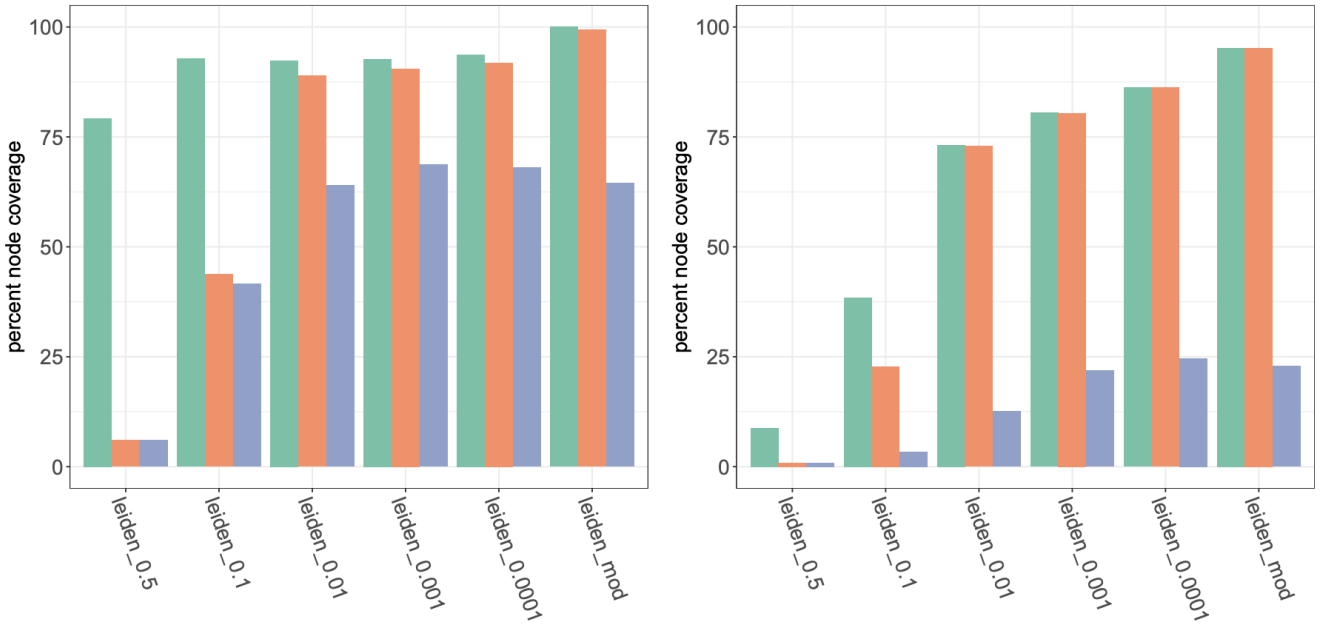
Clustering Accuracy



Results for ARI accuracy on LFR networks.
Results for AMI and NMI are similar.

- Orkut not shown because LFR failed to complete
- Two CEN and one Wiki_topcats not shown because LFR produced too many disconnected ground truth clusters

CM reduces node coverage



(a) Open Citations

(b) CEN

Figure 4: *Reduction in node coverage after CM treatment of Leiden clusters.* The Open Citations (left panel) and CEN (right panel) networks were clustered using the Leiden algorithm under CPM at five different resolution values or modularity. Node coverage (defined as the percentage of nodes in cluster of size at least 2) was computed for Leiden clusters • (lime green), Leiden clusters with trees and/or clusters of size 10 or less filtered out • (soft orange), and after CM treatment of filtered clusters • (desaturated blue).

- Green: original clustering
- Orange: after removing trees & small clusters
- Blue: after CM pipeline

Observations, part 1

- Leiden-CPM was the best of the tested methods (higher node coverage after CM treatment, and scalable to large networks)
- Leiden-Modularity is similar to Leiden-CPM with small resolution parameter values.

Observations, part 2

- Leiden-CPM depends on the resolution parameter value:
  - small values producing large node coverage but poorly connected clusters
  - large values producing small node coverage and small clusters that are generally well-connected
- So: trade-off between edge-connectivity and node coverage
- After CM, node coverage is substantially reduced

Observations, part 3

We noted:

- CM improves accuracy on LFR networks for Leiden-CPM and Leiden-Modularity, suggesting that both methods might be over-clustering.
- CM produces a drop in node coverage that can be large (especially for CPM, if the resolution parameter is small).

Observations, part 3 and Question

We noted:

- CM improves accuracy on LFR networks for Leiden-CPM and Leiden-Modularity, suggesting that both methods might be over-clustering.
- CM produces a drop in node coverage that can be large (especially for CPM, if the resolution parameter is small).

Perhaps these networks are not fully covered by communities?

Stochastic Block Models

- Popular generative model for networks with community structure
- SBMs in wide use using graph-tool by Peixoto
- We study them for clustering real-world and ground-truth clusters, and for generating synthetic networks with communities

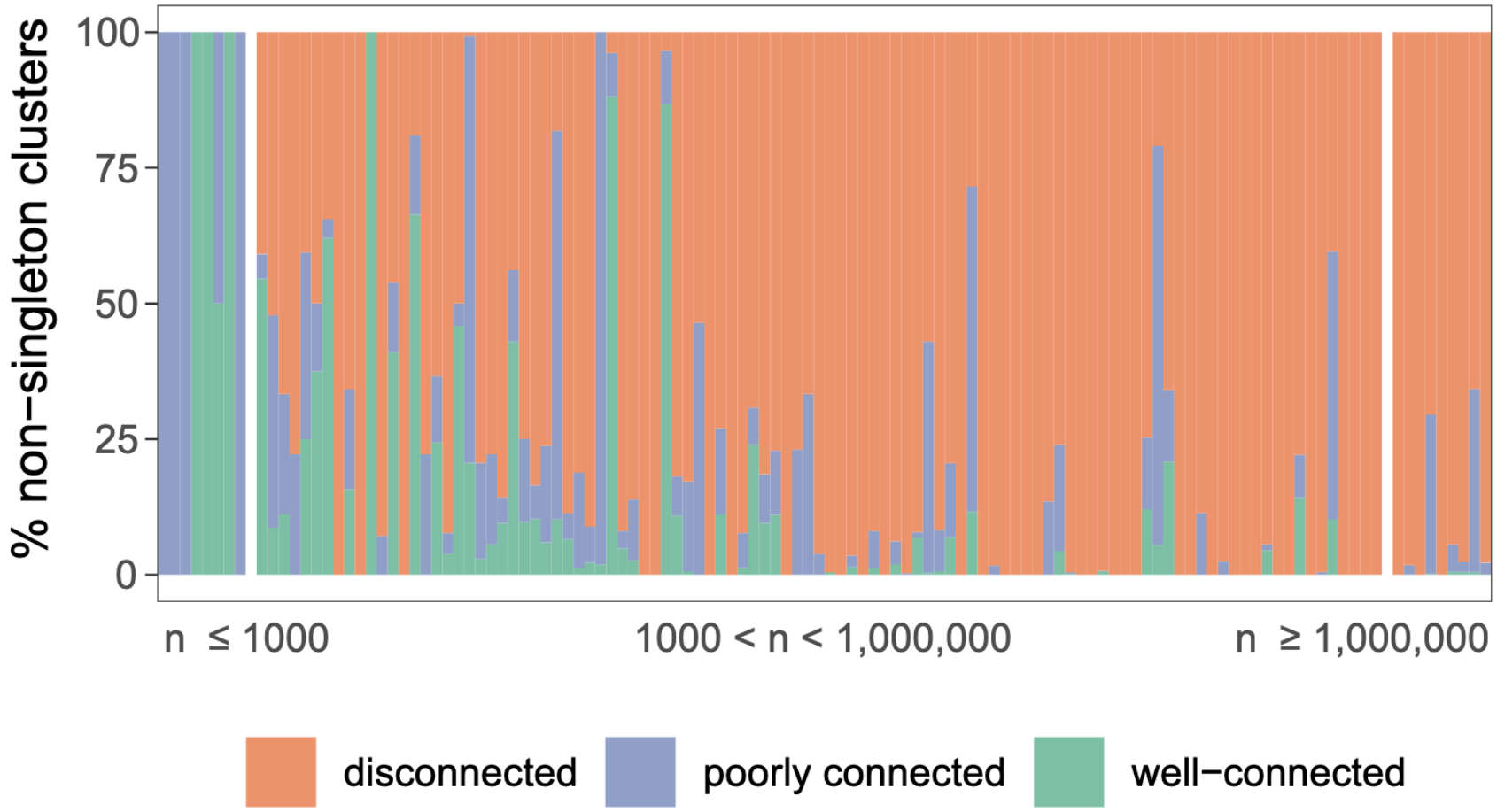
III. Clustering using Stochastic Block Models

"Clustering using Stochastic Block Models"

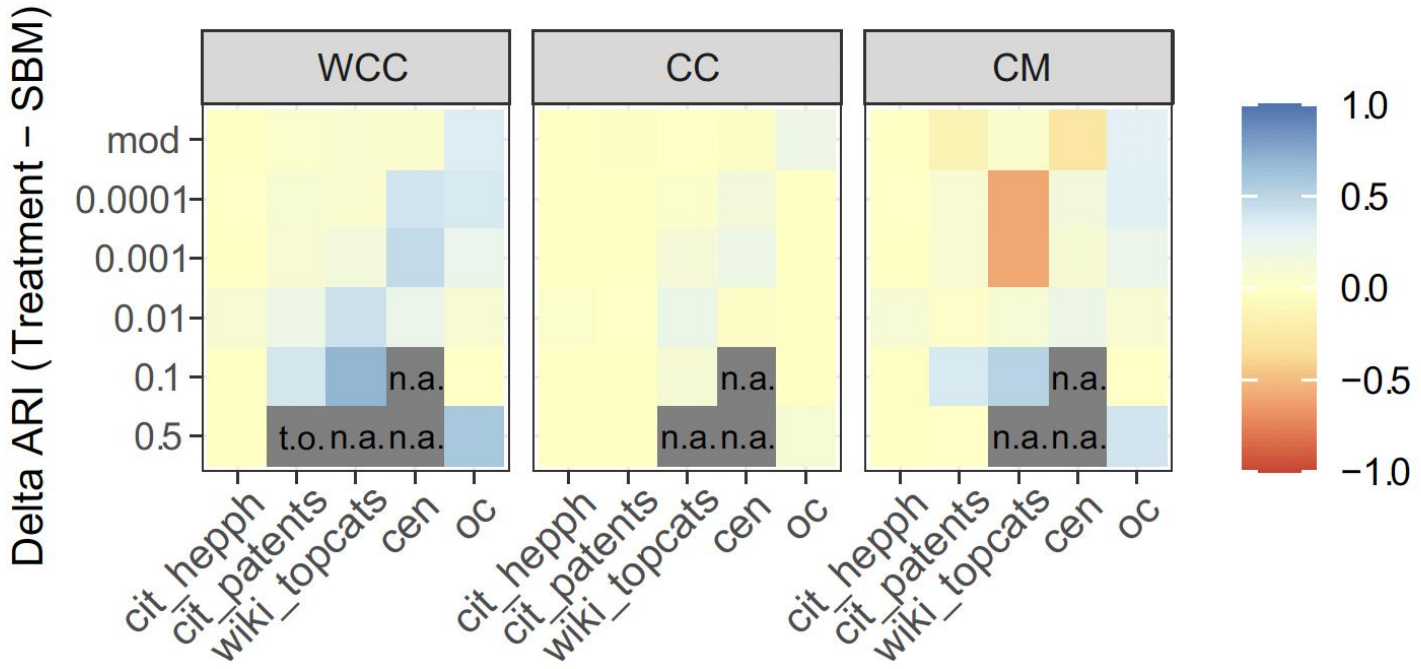
Park et al., Complex Networks and their Applications 2024

- Evaluated clustering using SBM for connectivity on 120 real-world networks
- Examined impact on accuracy on LFR networks (from CM paper) using three treatments
  - CM (Connectivity modifier) – but without filtering for size
  - Well-Connected Clusters (WCC) and
  - Connected Components (CC)

SBM clustering of 120 real-world networks



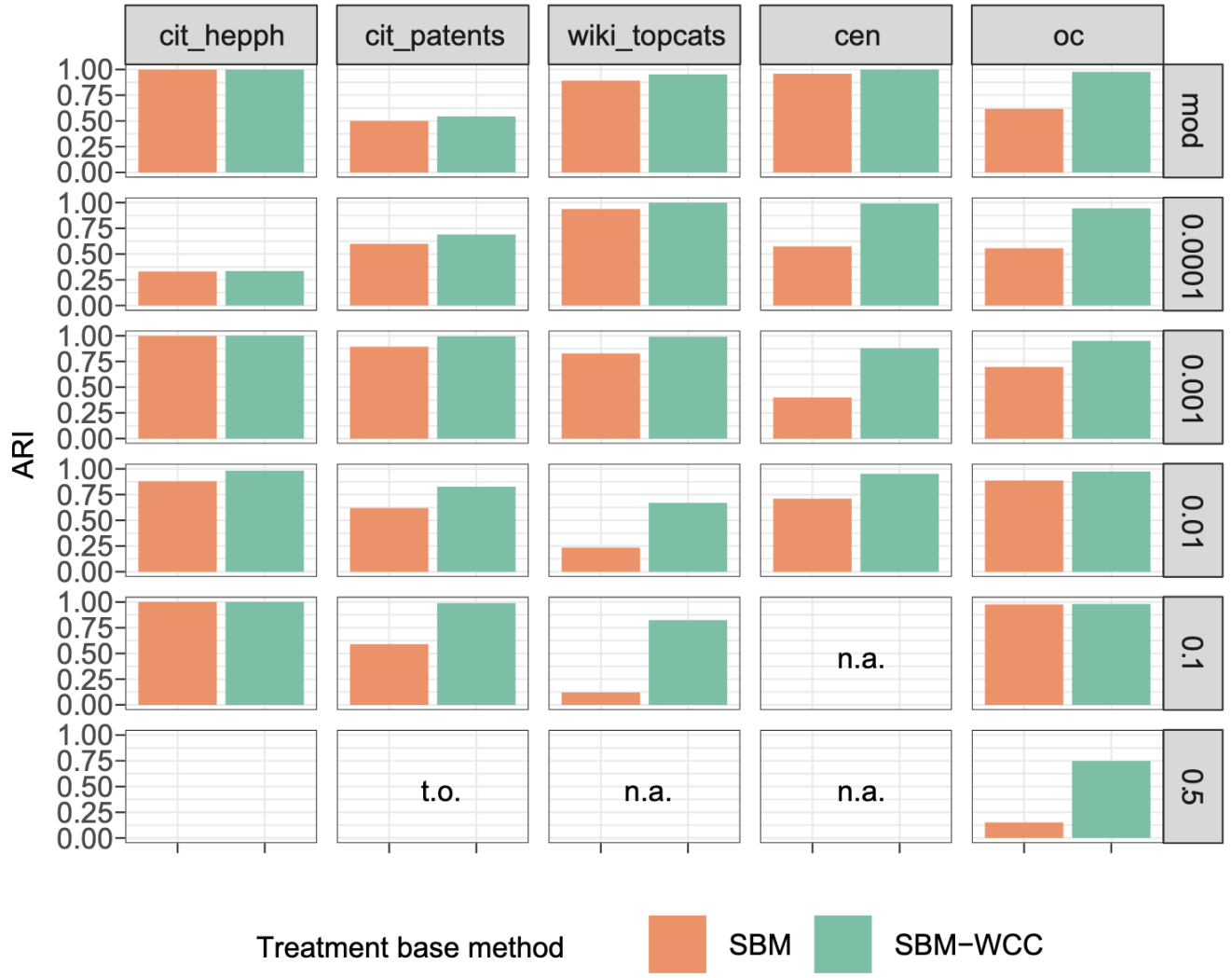
Impact on accuracy for LFR networks



- CM: Connectivity Modifier
- WCC: Well-connected clusters
- CC: Connected Components

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 net-

Impact of WCC (Well-Connected Clusters)



- ”n.a.” means no LFR network (or LFR network had too many disconnected ground truth clusters)
- One time-out for WCC
- In general, WCC improves accuracy

Take home points

- All tested clustering methods produced clusters that had small edge cuts.
- Two possible explanations:
  - Optimization problems in clustering lead to over-clustering
  - Not all of the network is occupied by valid communities.

Take home points

- All tested clustering methods produced clusters that had small edge cuts.
- Two possible explanations:
  - Optimization problems in clustering lead to over-clustering
  - Not all of the network is occupied by valid communities.
- Hence:
  - Simple techniques (WCC and CM) can improve accuracy of clusterings
  - Clusters should be checked for edge connectivity.
  - Ensuring edge-connectivity should be part of community detection methods

IV. Synthetic networks using SBMs

- Synthetic networks using graph-tool software for Stochastic Block Models
- Vaca-Ramirez and Peixoto evaluated fit to real-world network properties, not to clusterings
- In Anne et al., CNA 2024, we examine how well SBMs fit clustering properties

PHYSICAL REVIEW E
covering statistical, nonlinear, biological, and soft matter physics

Systematic assessment of the quality of fit of the stochastic block model for empirical networks

Felipe Vaca-Ramírez and Tiago P. Peixoto
Phys. Rev. E **105**, 054311 – Published 18 May 2022

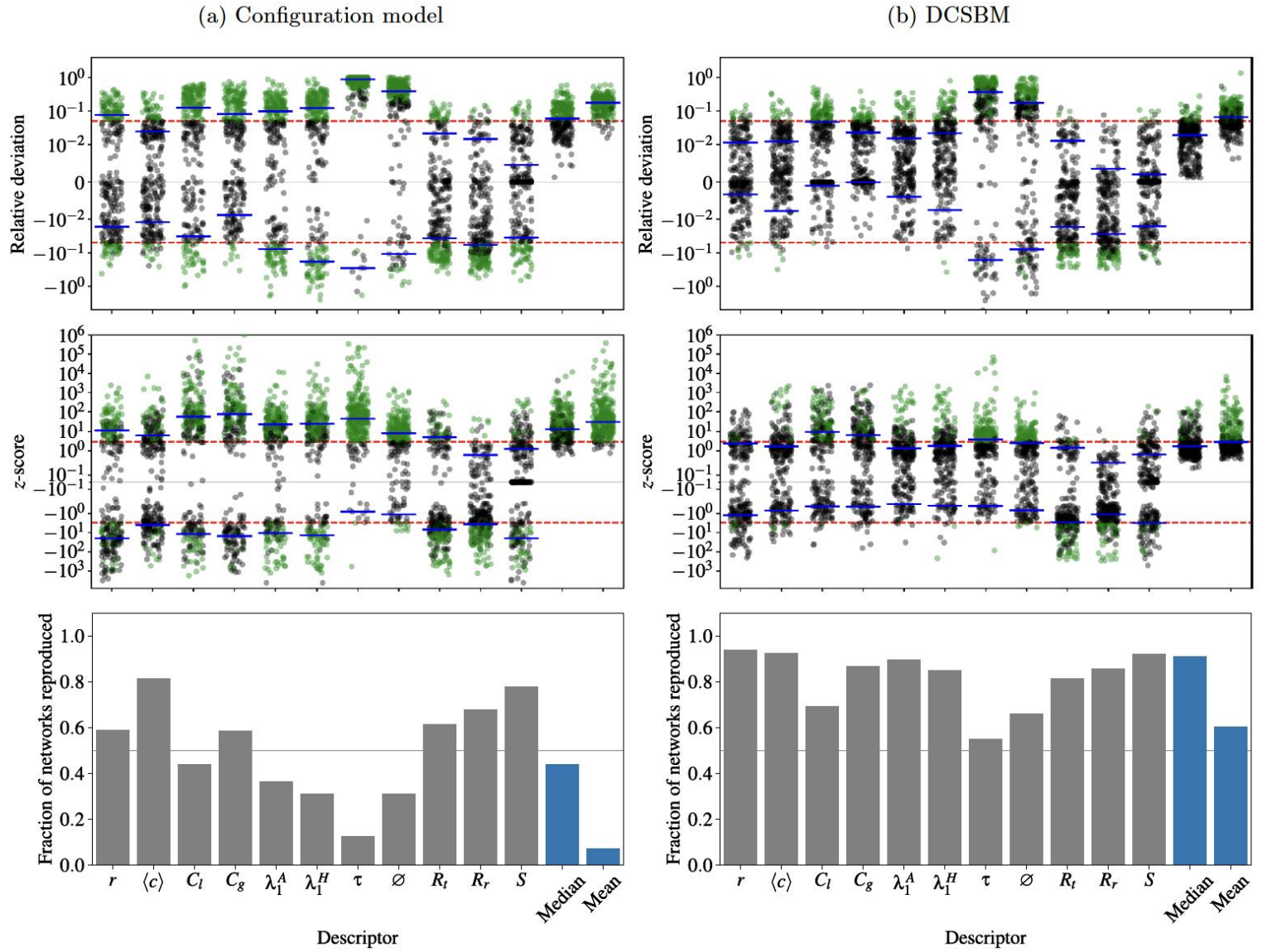


FIG. 3. Distribution of relative deviation (top), z score (middle), and fraction of networks reproduced (bottom) for (a) the configuration model and (b) the DCSBM, according to their respective predictive posterior distributions for each descriptor. We also show the median and

Vaca-Ramirez and Peixoto show that Degree-Corrected SBM has better fit to empirical properties (clustering coefficient, diameter, etc.) of real-world networks than the Configuration Model

Accuracy of cluster properties was not considered

SBM synthetic networks have disconnected clusters

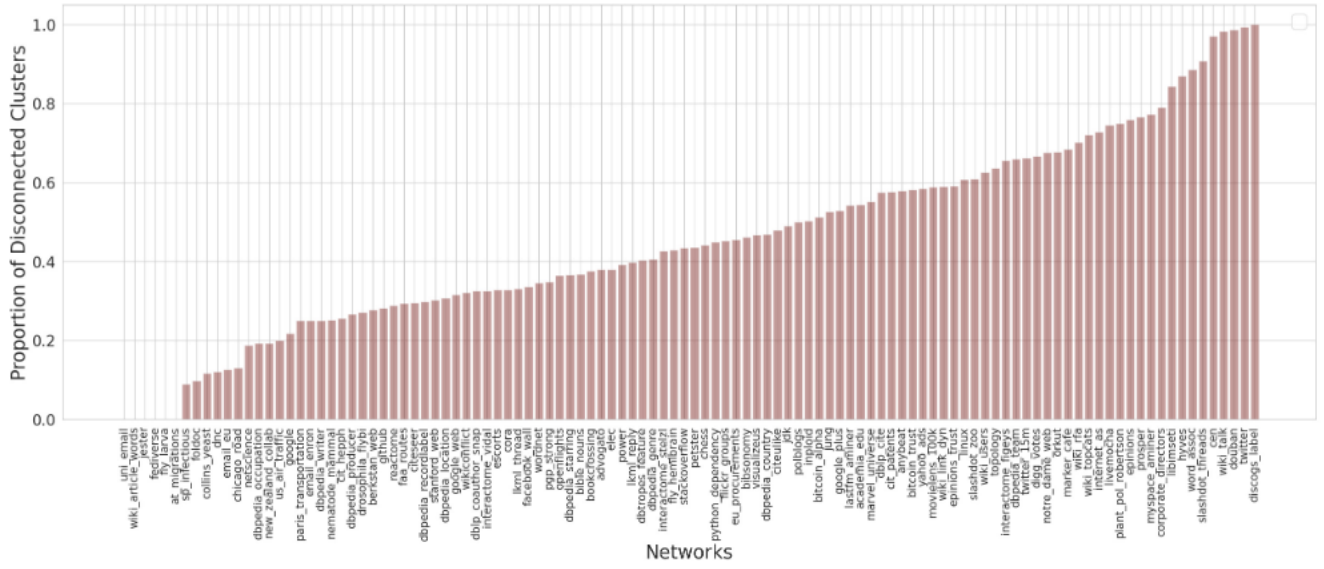


Fig. 2. Proportion of disconnected clusters in SBM generated networks. The x-axis shows 110 SBM networks generated using parameters from real world networks clustered with the Leiden+CM (Connectivity Modifier) pipeline (training data). The SBM method failed to reproduce the guaranteed connectivity of Leiden+CM clusters.

Anne et al. CNA 2024: Using parameters from clustered real-world networks, we generated Degree Corrected SBMs using graph-tool.

These networks had many disconnected ”ground truth” clusters.

RECCS: improving fit to input clustering statistics

Given input parameters from a clustered real-world network, RECCS

- Generates a degree-corrected SBM
- Removes self-loops and excess parallel edges
- Adds edges to achieve connectivity for each cluster
- Adds edges to improve fit to node degree sequence
- Finally adds in the “outliers”

RECCS simulator: Builds on SBM generation

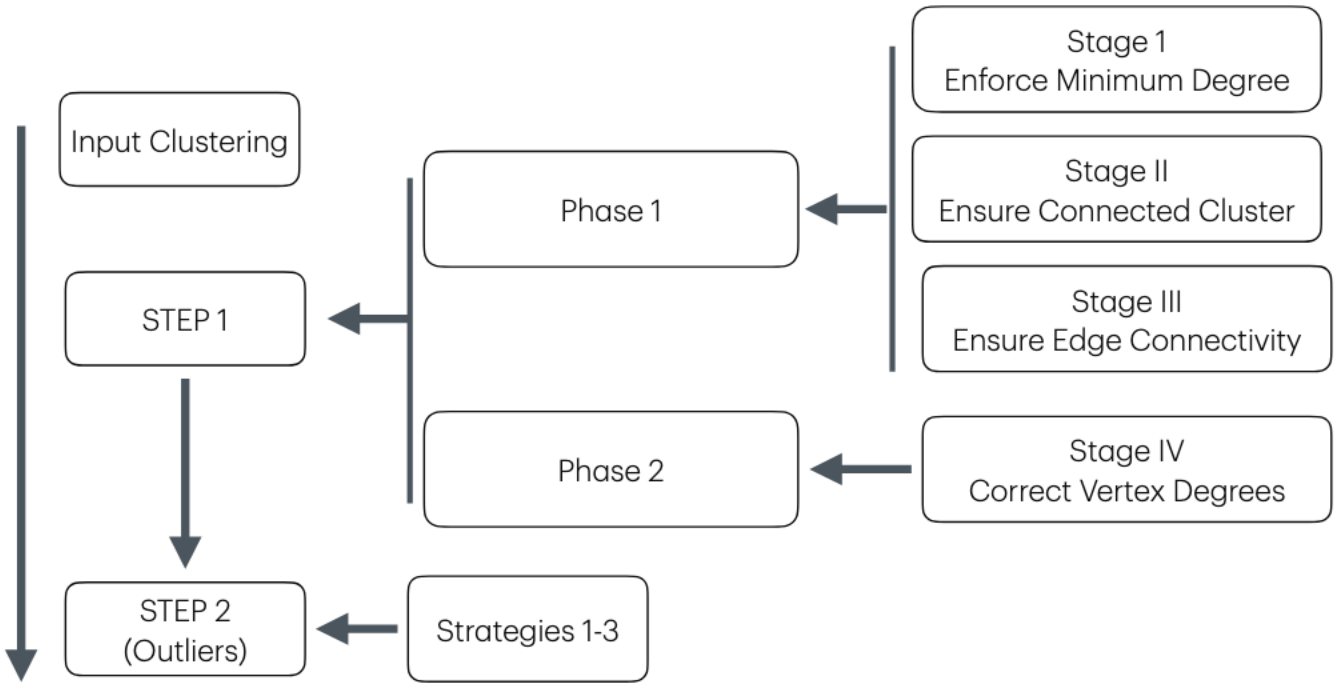


Fig. 1. RECCS **Workflow.** The input to RECCS is a collection of parameters computed on a real-world network N and an estimated clustering. Step 1 of RECCS pro-

RECCS v1 and v2 differ in Stage IV (v2 uses more information)

Stragies I-3 for how to add outliers also differ, with Strategy 1 just using SBMs (every outlier is in its own cluster)

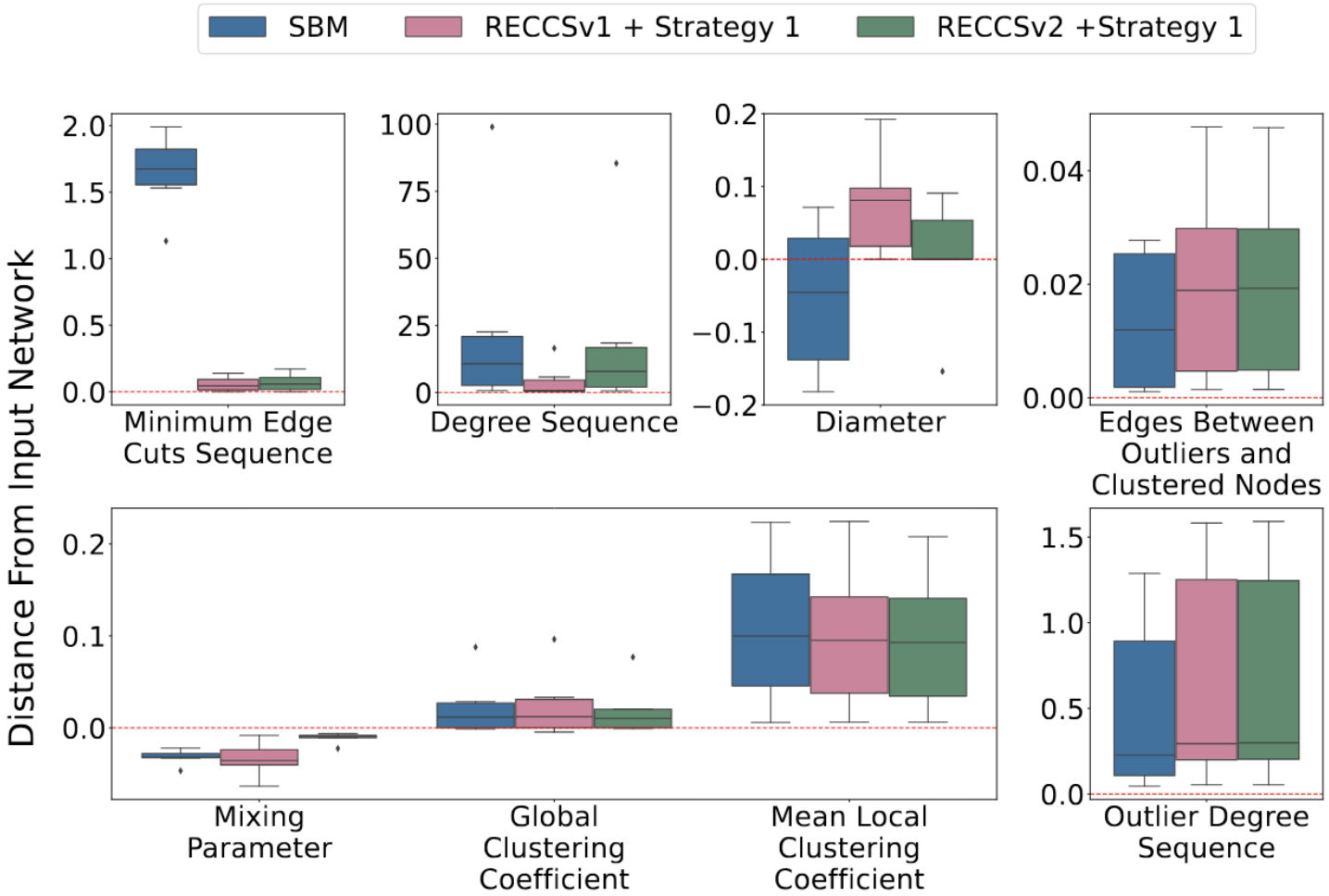


Fig. 3. Comparing SBM to the RECCS pipelines on the test networks using Leiden-CPM(0.01)+CM. We compare SBM networks to networks produced using

Results on six real-world networks clustered using Leiden-CPM(0.01)+CM.

Results on other clusterings of these networks show the same trends.

Distance is positive when s>s', where

- s = statistic for real-world network
- s' = statistic for the synthetic network

Overall summary

- Cluster connectivity is important for clustering and synthetic network generation
- Simple ad hoc techniques are helpful
- Rigorous mathematical approaches and models are needed

Our codes are open source

- CM
  - enables integration of new clustering methods
  - algorithmic parameters (e.g., what “well-connected” means) are user-defined
  - has been run on networks with 75M nodes
- WCC and RECCS are slower, but under development.
- We welcome collaborations.
- See https://github.com/illinois-or-research-analytics/cm_pipeline and https://github.com/illinois-or-research-analytics/lanne2_networks
- See https://tandy.cs.illinois.edu/bibliometrics.html for papers

Acknowledgments

- Connectivity Modifier: Minhyuk Park, Yasamin Tabatabaee, Vikram Ramavarapu, Baqiao Liu, Vidya Kamath Pailodi, Rajiv Ramachandran, Dmitriy Korobskiy, Fabio Ayres, Geoge Chacko,
- Using SBMs in clustering: Minhyuk Park, Daniel Feng, Siya Digra, The-Anh Vu-Le, and George Chacko
- RECCS: Lahari Anne, The-Anh Vu-Le, Minhyuk Park, and George Chacko
- Supported by the Insper-Illinois Partnership, Oracle, Digital Science, Google, and the Grainger Foundation
- Graduate students
- Undergraduate students