The Problem of Cluster Connectivity in Community Detection Methods

Tandy Warnow

Siebel School of Computing and Data Science

University of Illinois Urbana-Champaign

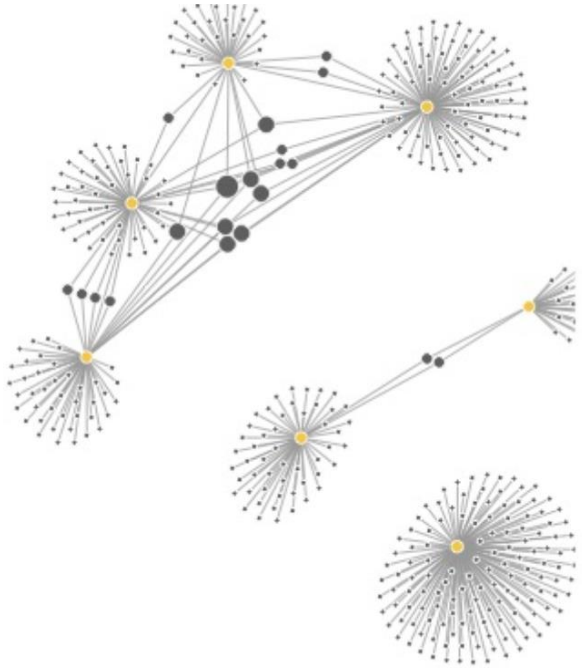
(a long time ago, in 1423 SNL-Albuquerque)

Supported by the National Science Foundation, the Illinois-Illinois Partnership, Oracle, Digital Science, 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

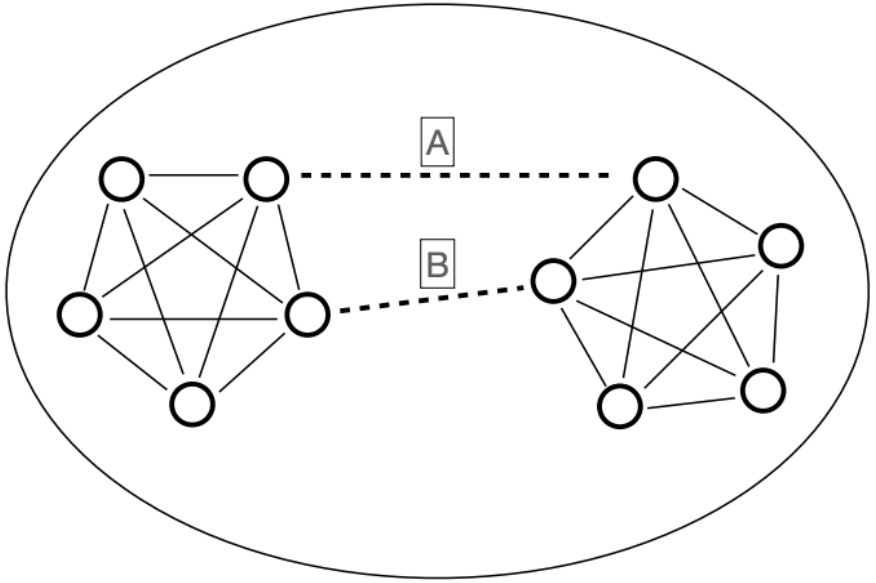
I. Cluster connectivity: what this means, what is known
II. The Connectivity Modifier (Park et al. PLOS Complex Systems 2024) – results on Leiden, IKC, MCL, and Infomap
III. Clustering using Stochastic Block Models (Park et al., Complex Networks and their Applications 2024)
IV. Synthetic networks using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

Overall summary

- Cluster connectivity is important for clustering and synthetic network generation
- Yet – standard methods fail to produce well-connected clusters
  - And some methods produce disconnected clusters!
- Simple ad hoc techniques are helpful in ameliorating these problems

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 disconnects 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

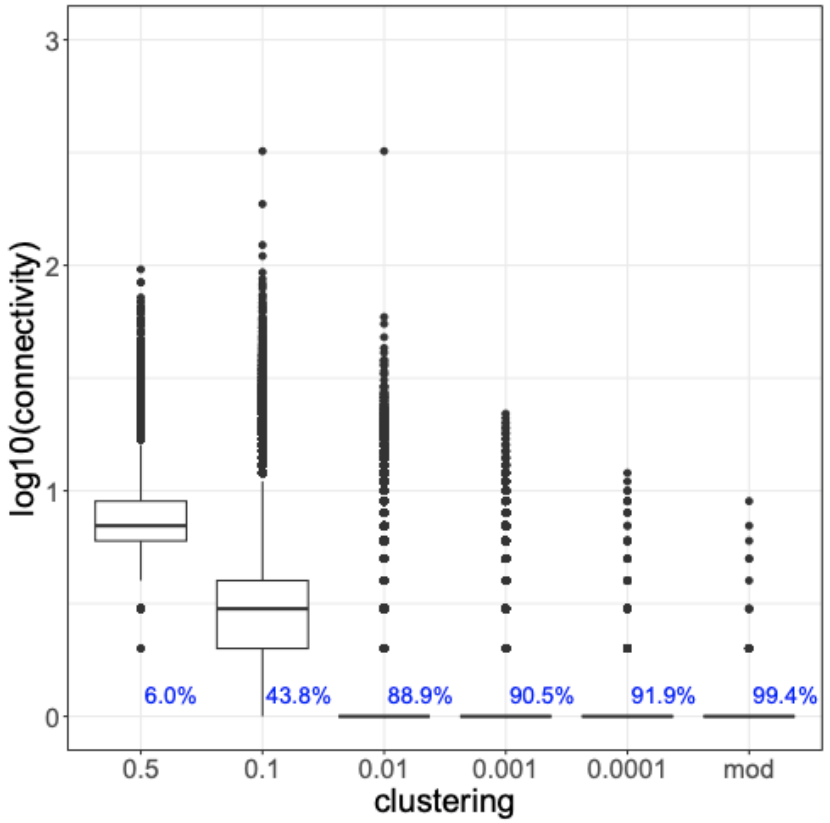
The Constant Potts Model (CPM) optimization score depends on the resolution parameter γ

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

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 indicates node coverage

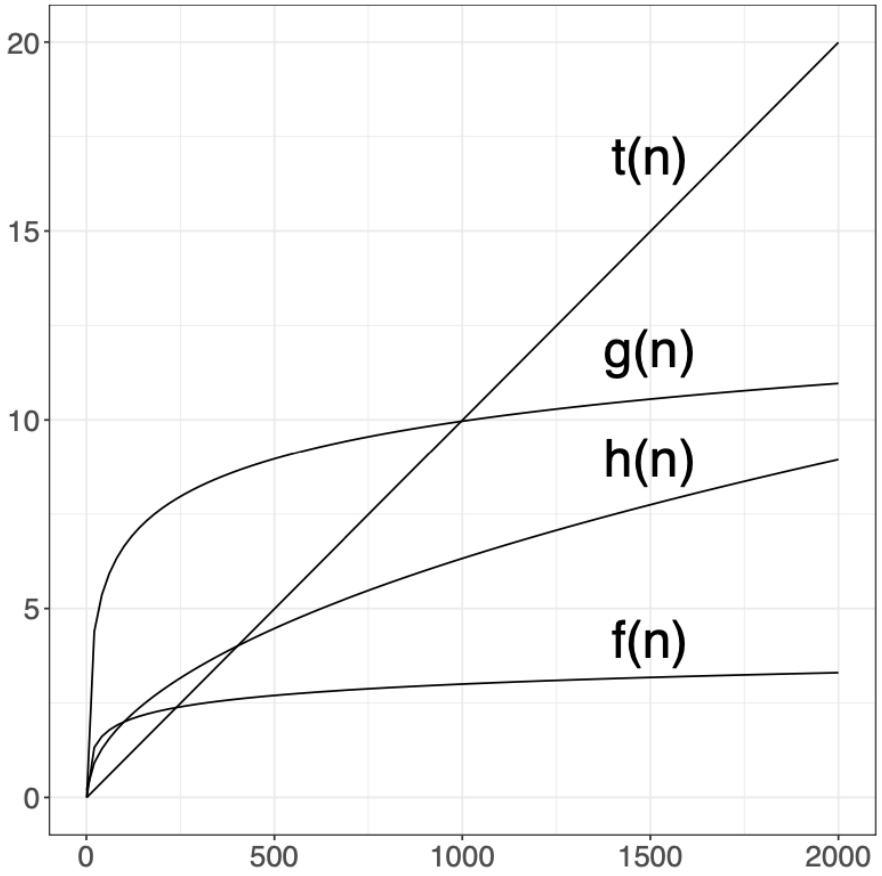
Trade-off between node coverage and edge-connectivity

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

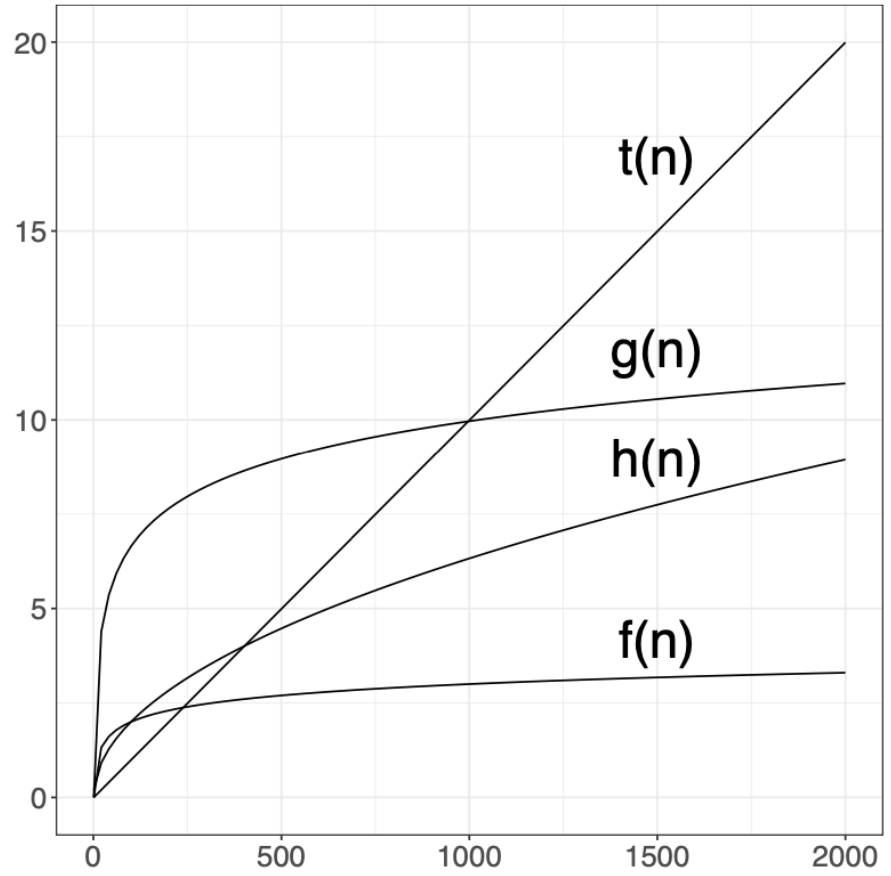
$f(n) = \log_{10} n$

$g(n) = \log_2 n$

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

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 LFR synthetic networks based on these networks.

Only Leiden and IKC completed on Open Citations.

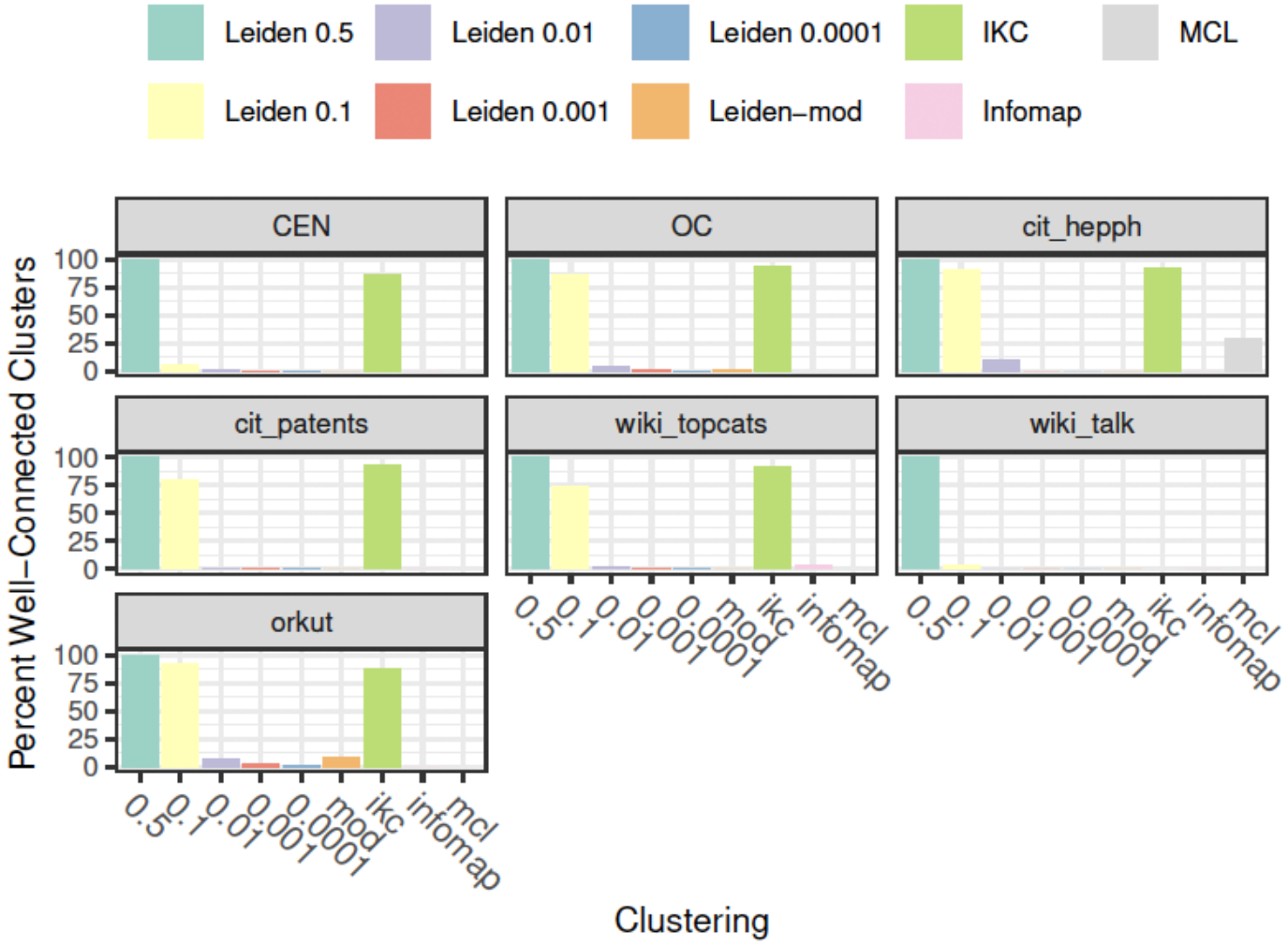
IKC had very low node coverage

Community Detection Methods:

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

LFR : Lancichinetti, Fortunato, and Radicchi. 2008. Benchmark graphs for testing community detection algorithms. Physical Review E 78, 4 (Oct. 2008), 046110.

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

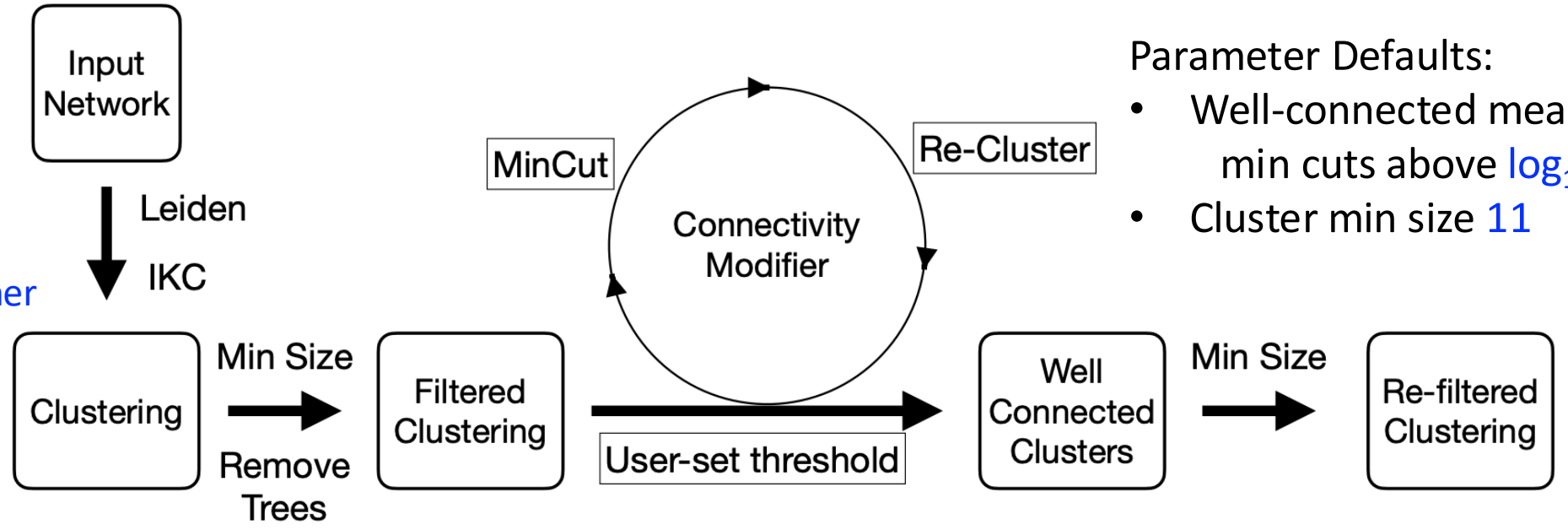
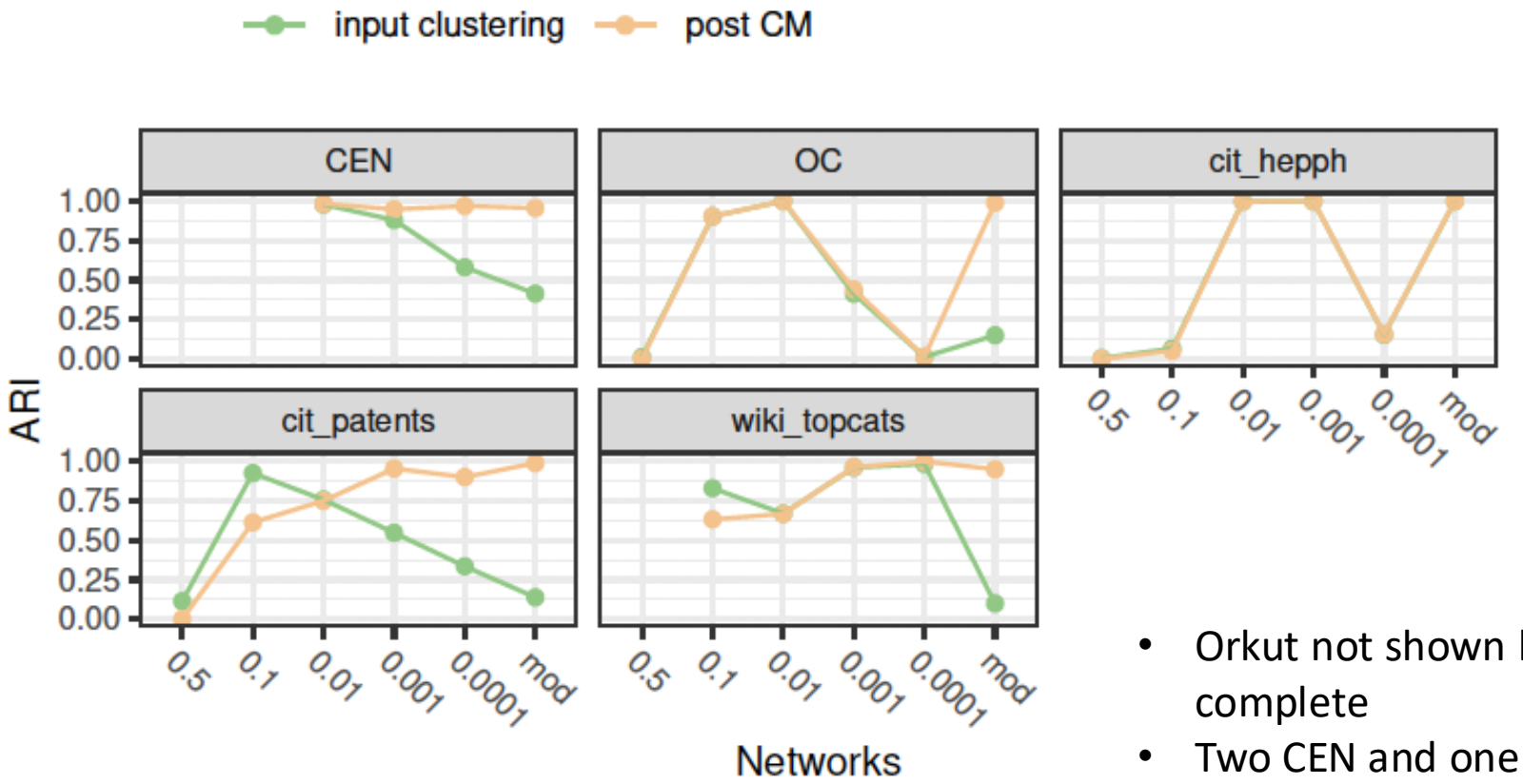


Figure 3: *Connectivity Modifier Pipeline Schematic* The four-stage pipeline depends on user-

CM improves accuracy on synthetic networks

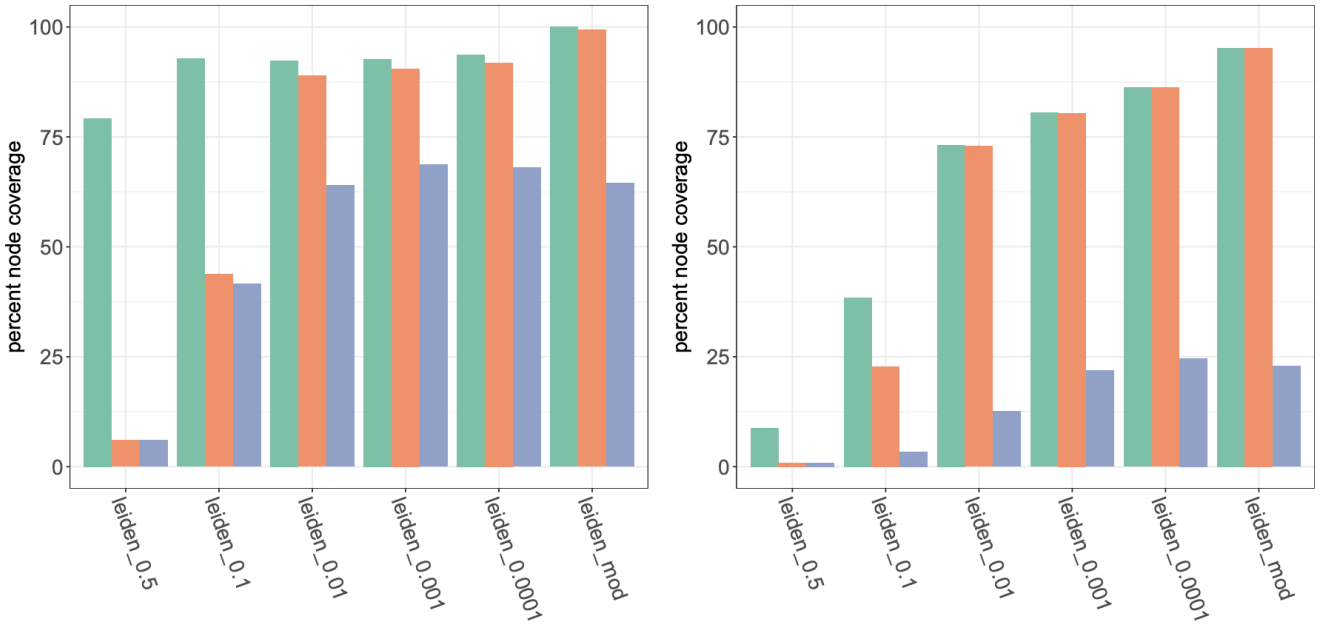
Clustering Accuracy



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

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 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

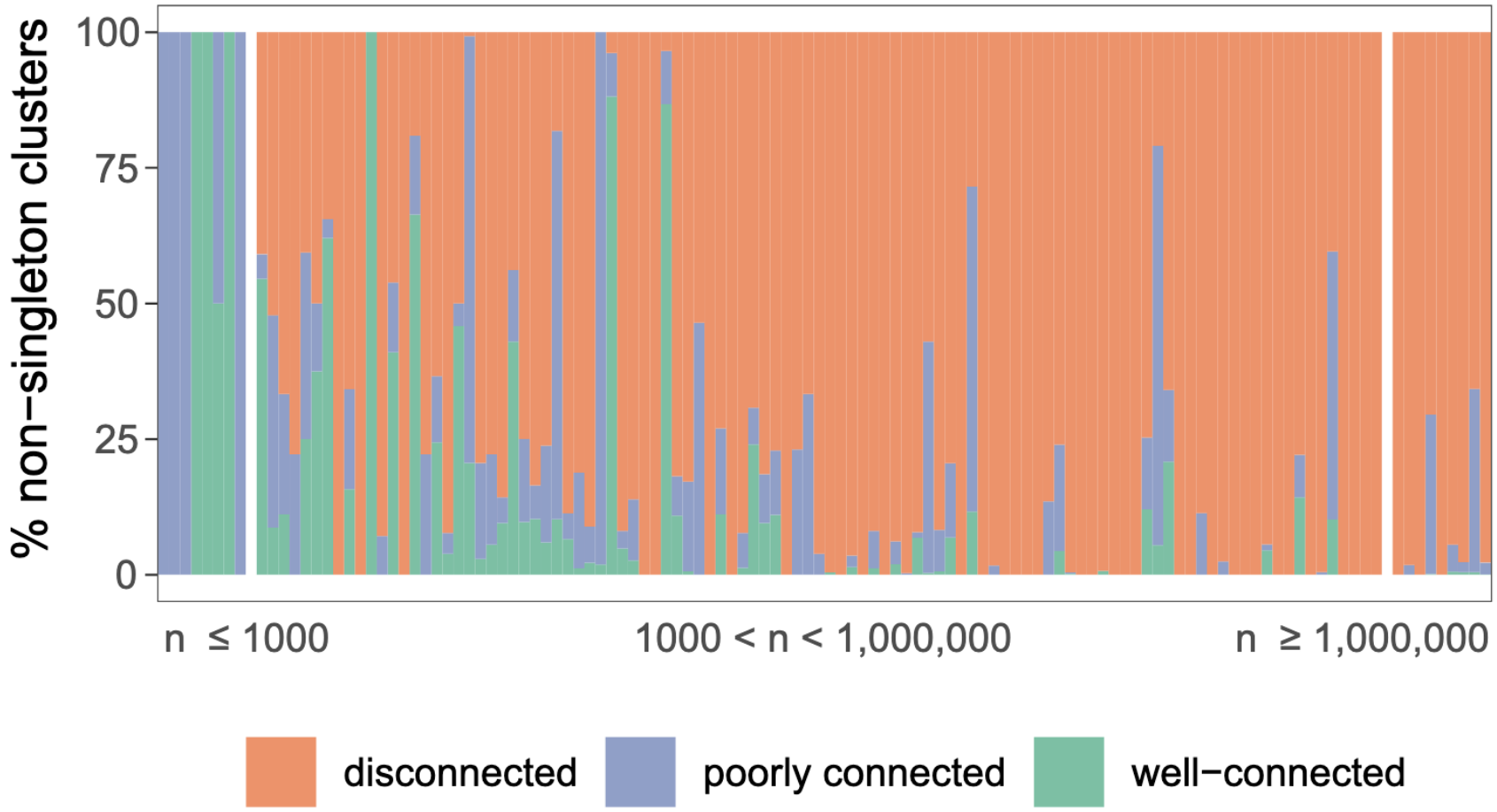
- Popular generative model for networks with community structure
- SBMs in wide use using graph-tool by Peixoto
- We study them for clustering real-world networks, and for generating synthetic networks with communities

III. Clustering using Stochastic Block Models

"Improved Community Detection using Stochastic Block Models"

Park et al., Complex Networks and their Applications 2024 (under review in PLOS Complex Systems)

SBM clustering of 120 real-world networks



Why do SBMs produce disconnected clusters?

- SBM clustering in graph-tool (Peixoto) obtained by optimizing for the description length
- We examine the formula for the degree-corrected SBM

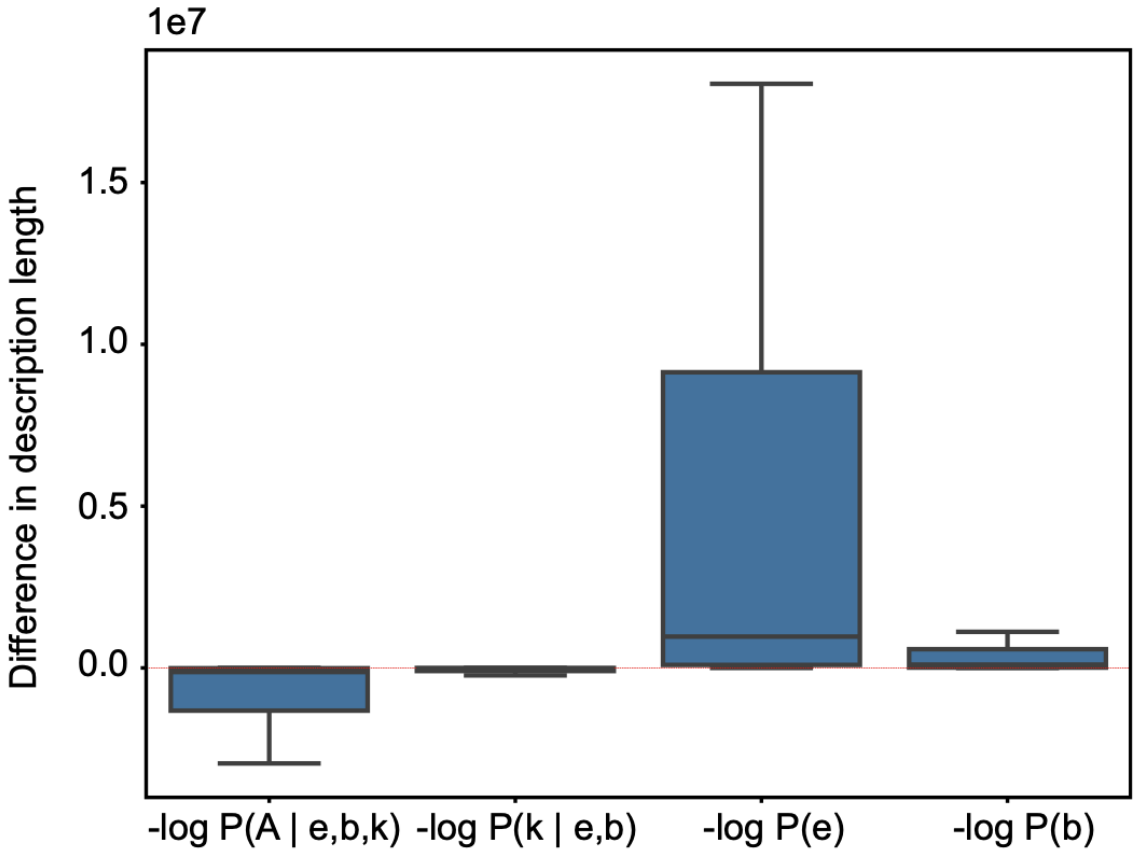
Description Length (DL) formula for degree-corrected SBM

- A be the adjacency matrix, defined by the network N,
- b be the block (cluster) assignment, which represents the clustering of N,
- k be the degree vector, defined by A,
- e be the edge count matrix, defined by A and b.

Eq (1) provides the formula for the description length $\mathrm{DL}(A, b)$ of a network A and a clustering b under the Degree Corrected (DC) model:

$$\mathrm{DL}(A, b) = -\log p(A|b, e, k) - \log p(k|b, e) - \log p(b) - \log p(e) \tag{1}$$

Impact of CC (reducing to connected components)



Results shown for 71 networks where degree-corrected SBM selected

We show: DL(SBM+CC) – DL(SBM)

Positive values favor not splitting disconnected clusters

Total of these values is positive because of –log p(e)

Fig 8. SBM(DC)+CC to SBM(DC) difference for description length components. We show the contribution of $-\log p(e)$ term on the description lengths

Why SBMs favor disconnected clusters

$$-\log p(e) = \log \binom{B(B+1)/2 + E - 1}{E}$$

B is the number of blocks (clusters)
E is the number of edges (fixed)

Hence: a small number of blocks is preferred

Addressing SBM clustering: CC, WCC, and CM

- Connected Components (CC): return connected components
- Connectivity modifier (CM) without filtering for size
- Well-Connected Clusters (WCC): CM without re-clustering

Well-Connected Clusters (WCC): CM without reclustering

Note: WCC DOES NOT RECLUSTER

Parameter Defaults:

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

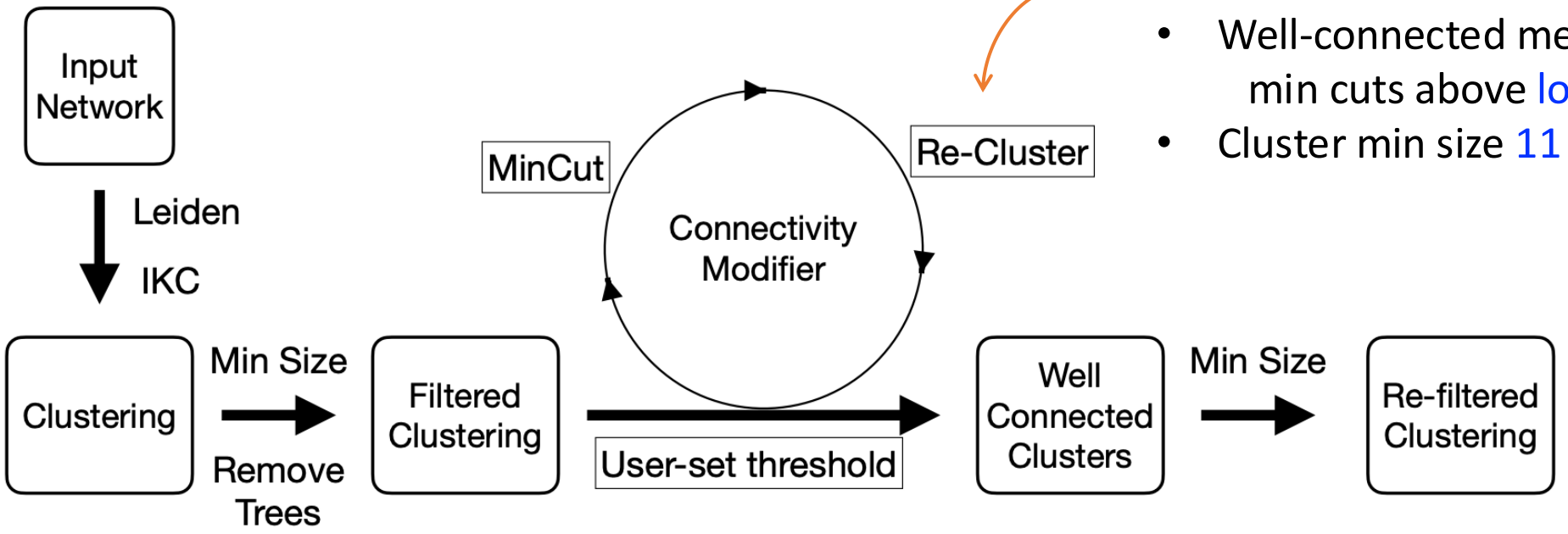
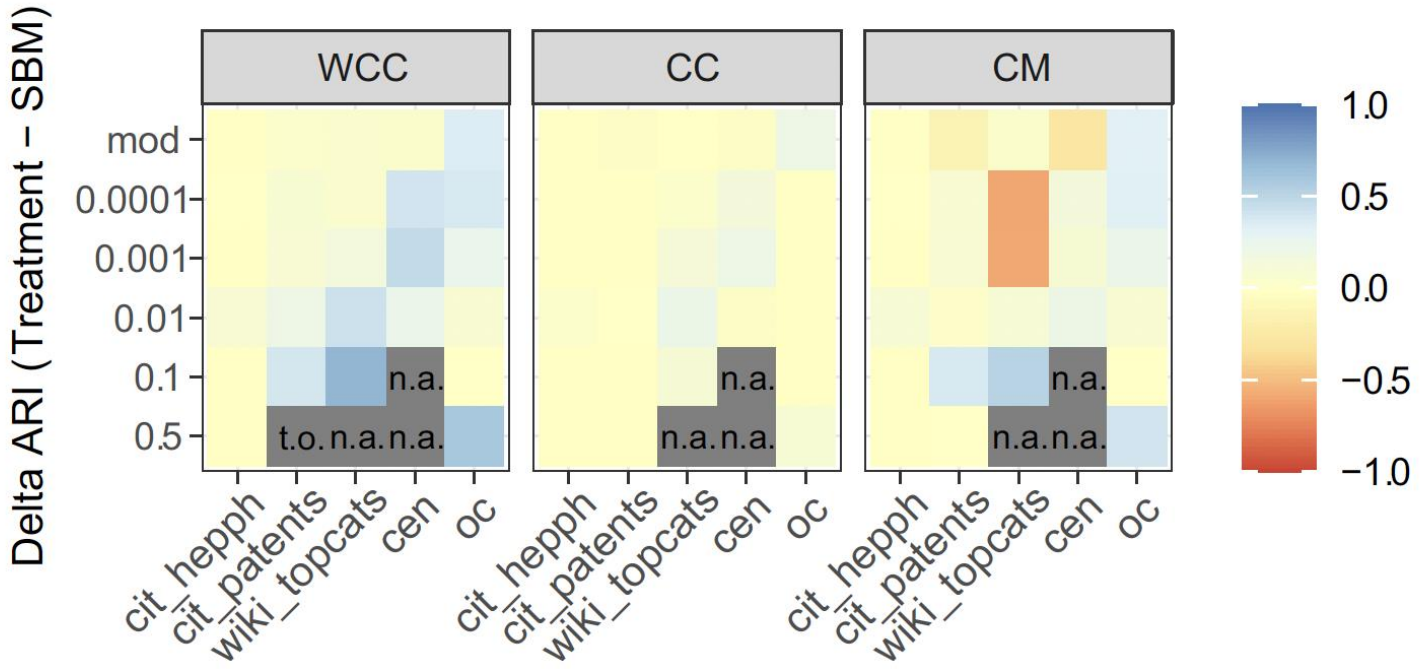


Figure 3: *Connectivity Modifier Pipeline Schematic.* The four-stage pipeline depends on user-

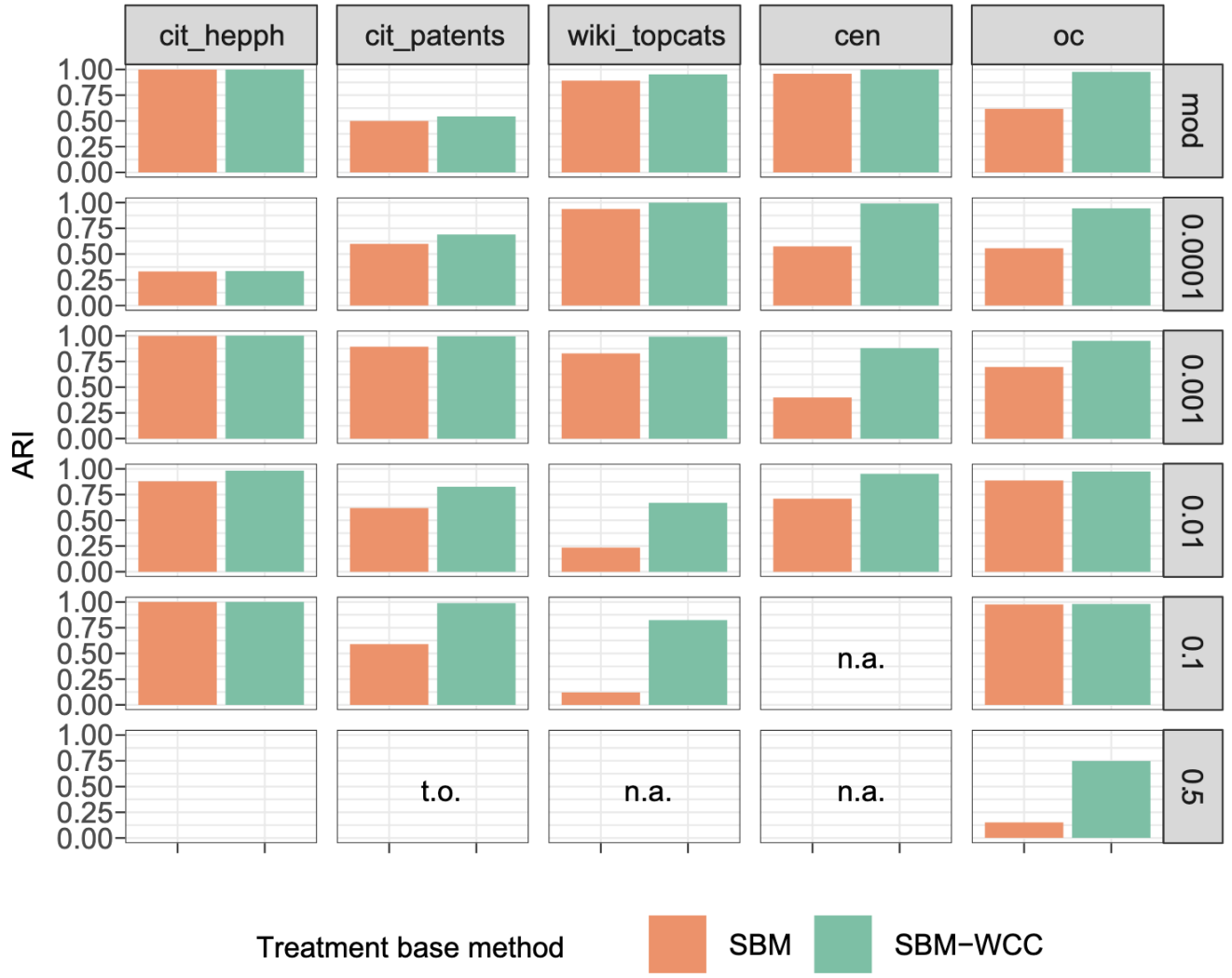
Impact on clustering accuracy on synthetic LFR networks



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

Fig. 3. Experiment 3: Impact of Treatment on ARI Scores of Selected SBM (heatmap).

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 (so far)

- All tested clustering methods produced clusters that had small edge cuts.
- Post-processing to improve edge-connectivity can improve clustering accuracy

- Two possible explanations:
  - Optimization problems in clustering lead to over-clustering
  - Not all of the network is occupied by valid communities.

IV. Synthetic networks using SBMs

- Vaca-Ramirez and Peixoto evaluated fit of SBMs (from graph-tool) to real-world network properties, not to clusterings
- We present two methods for synthetic networks that improve on SBMs:
  - RECCS (Anne et al., in press, Advances in Complex Systems)
  - EC-SBM (Vu-Le et al., in press, Applied Network Science)

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

Vaca-Ramirez and Peixoto (2022)

- Given a clustered real-world network, graph-tool SBM takes the following parameters
  - Node-to-cluster assignment
  - Number of edges within each cluster and between every two clusters
  - Degree of each node

- Given real-world networks, Vaca-Ramirez and Peixoto (2022):
  - computed degree-corrected SBM clusterings using graph-tool
  - gave the parameters from these clusterings to graph-tool
  - gave the degree sequence to the configuration model
  - compared the two synthetic networks w.r.t. network criteria

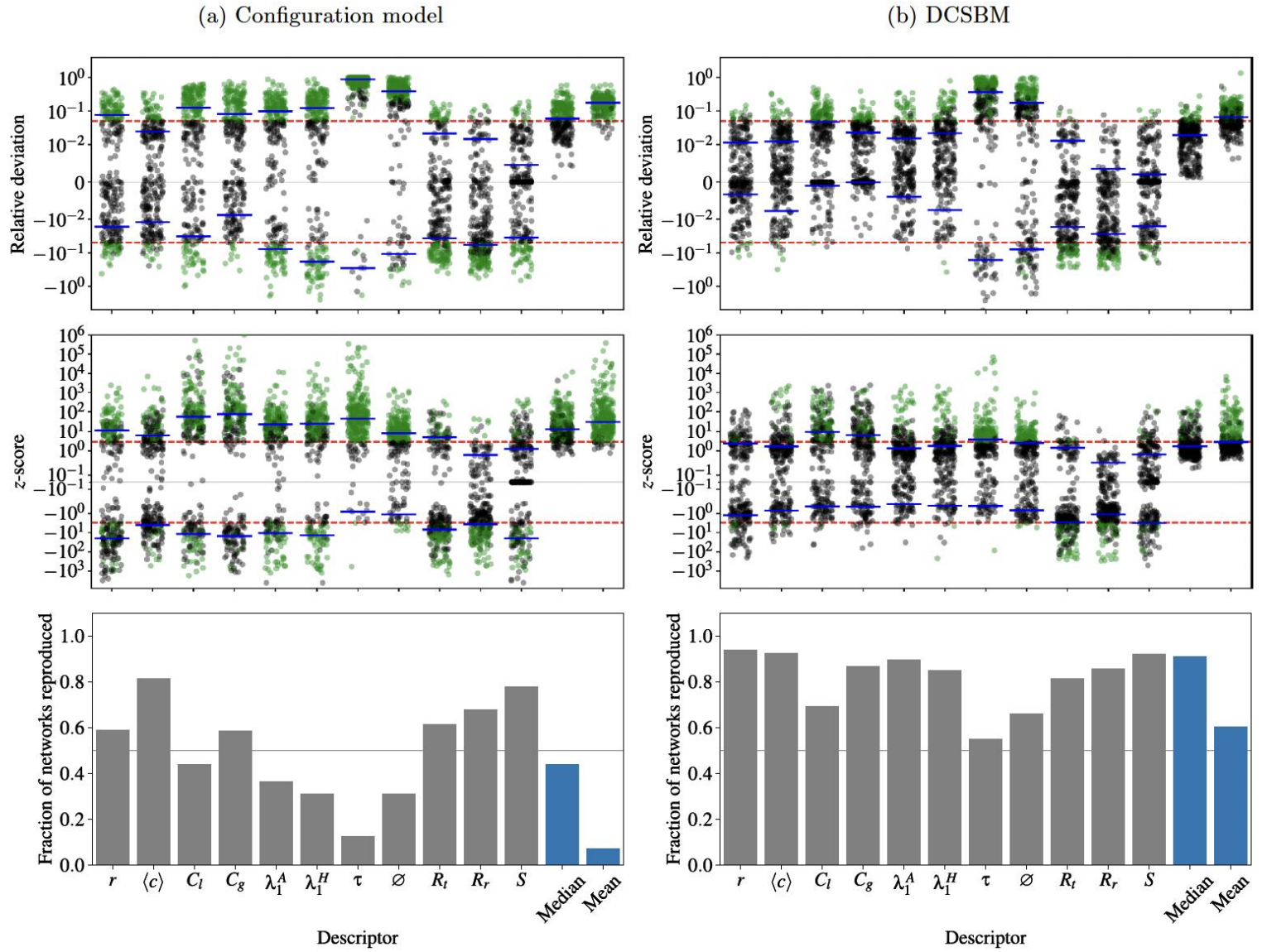


FIG. 3. Distribution of relative deviation (top), z score (middle), and fraction of networks reproduced (bottom) for (a) the configuration model and (b) the DCSBM, according to their respective predictive posterior distributions for each descriptor. We also show the median and

Results:

DCSBM better than Configuration Model for network properties (clustering coefficient, diameter, etc.) of real-world networks

Accuracy of cluster properties was not considered

Anne et al 2024: SBM synthetic networks have disconnected clusters

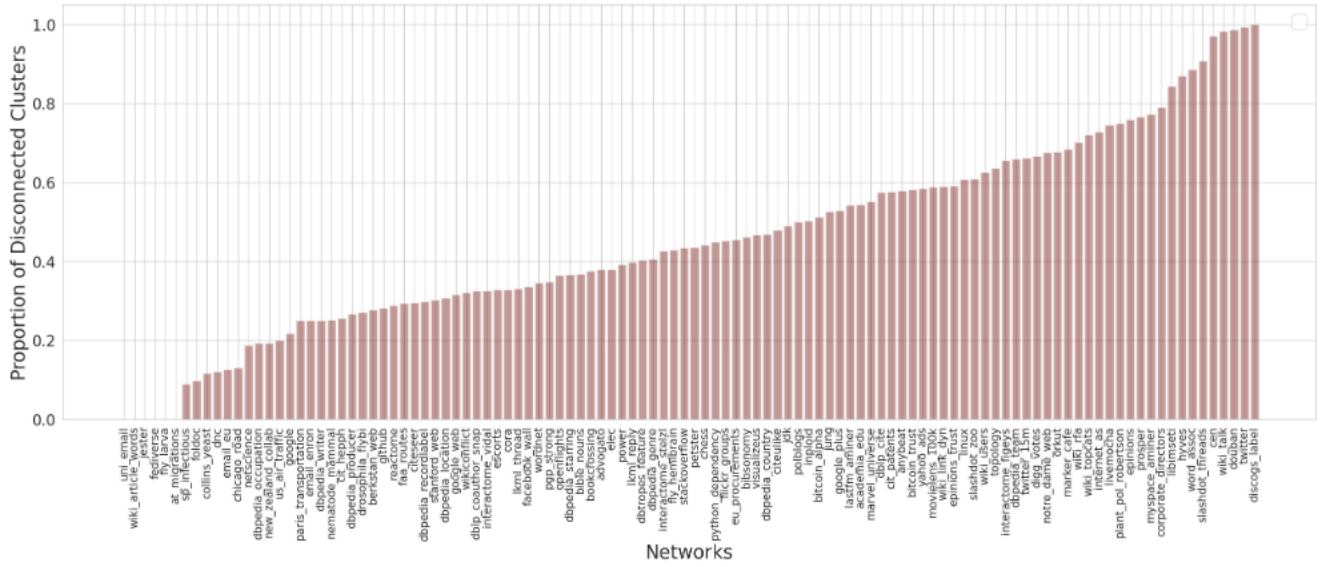


Fig. 2. Proportion of disconnected clusters in SBM generated networks. The x-axis shows 110 SBM networks generated using parameters from real world networks clustered with the Leiden+CM (Connectivity Modifier) pipeline (training data). The SBM method failed to reproduce the guaranteed connectivity of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

Anne et al. CNA 2024: Using parameters from Leiden+CM clusterings of real-world networks, we generated DC-SBMs using graph-tool.

- Ground-truth clusters are often disconnected
- Frequency increases with size of network

Vu-Le et al. 2025: Disconnected clusters occur for other clusterings

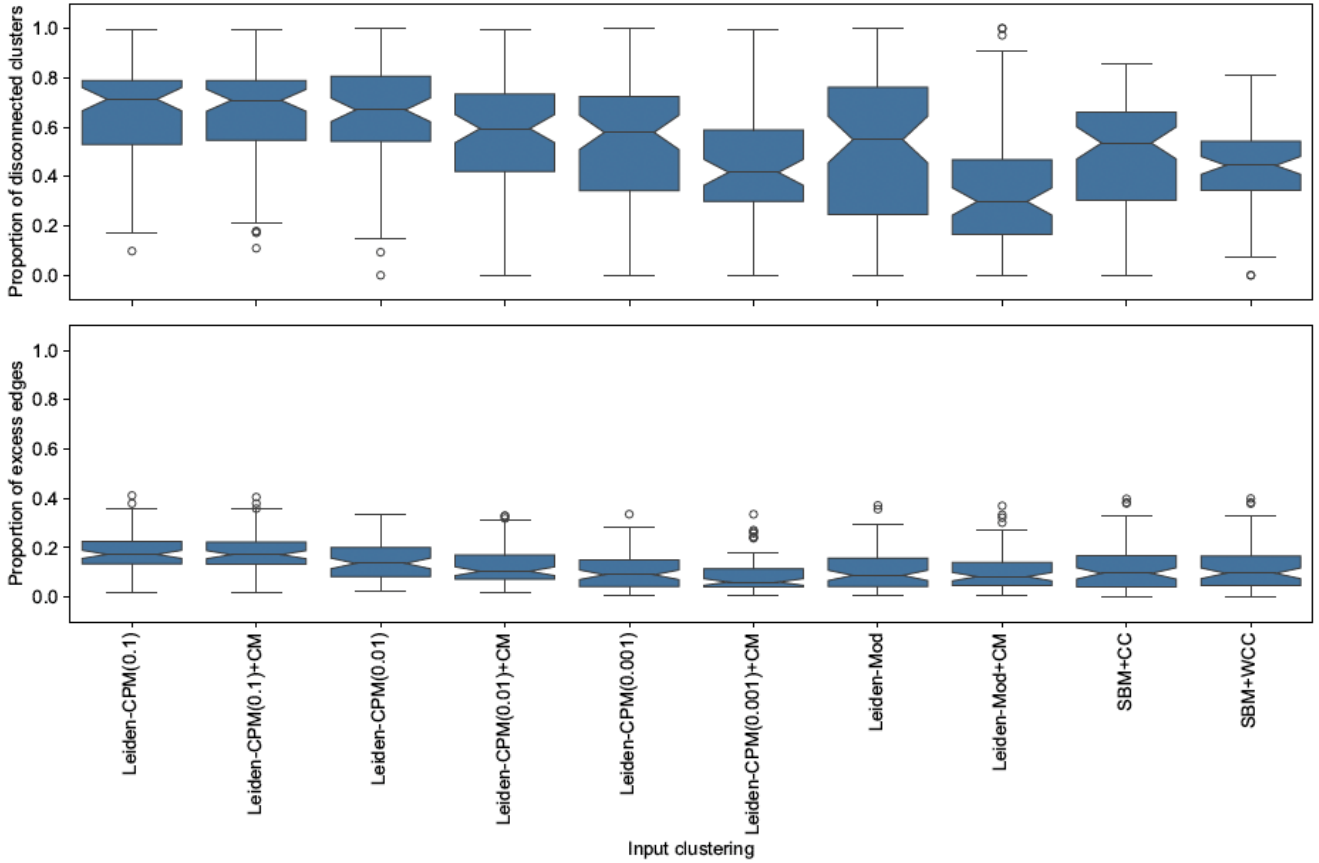


Fig. 1: Proportion of disconnected clusters (top) and excess edges (bottom) in synthetic networks generated by SBM. SBM is given network and clustering statistics for 74 networks, each clustered by one of the clustering methods specified on the horizontal axis, each of which is guaranteed to produce connected clusters.

Our question: if we cluster the real-world networks using other methods that are guaranteed to produce connected clusters, and give these parameters to DCSBM, are the clusters in these synthetic networks connected?

Observations: For all clustering methods:

- SBMs have many disconnected clusters
- SBMs have many excess edges (parallel edges and self-loops)

The problem

Problem: graph-tool SBMs produce disconnected ground truth clusters

Goal: New synthetic network generators that are more accurate.

Given parameters from a clustered real-world network, we want the synthetic network to:

- Match accuracy of graph-tool SBM for network properties
- Always have connected ground-truth clusters
- Come close to the same cluster edge-connectivity

Our synthetic network methods

- RECCS (Anne et al. 2025, in press Advances in Complex Systems)
- Edge-Connected SBM (Vu-Le, in press, Applied Network Science)

Both use synthetic networks produced by graph-tool SBMs

Both take advantage of graph-tool SBMs producing many excess edges

They use different algorithmic techniques

Edge-Connected SBM

- EC-SBM (Vu-Le et al., Applied Network Science, in press)
- Given a clustered real-world network, we define:
  - Clustered subnetwork: induced by nodes in non-singleton clusters
  - Unclustered subnetwork: induced by nodes in singleton clusters
- EC-SBM: Given parameters from a clustered real-world network N:
  - Construct synthetic clustered subnetwork N1
  - Create synthetic on the complement (N \ N1)
  - Merge the two networks

Edge-Connected SBM

- Uses the following parameters from a clustered real-world network:
  - Degree sequence
  - Assignment of nodes to clusters
  - Matrix of # edges within clusters and between clusters
  - Edge-connectivity within clusters

Edge-Connected SBM

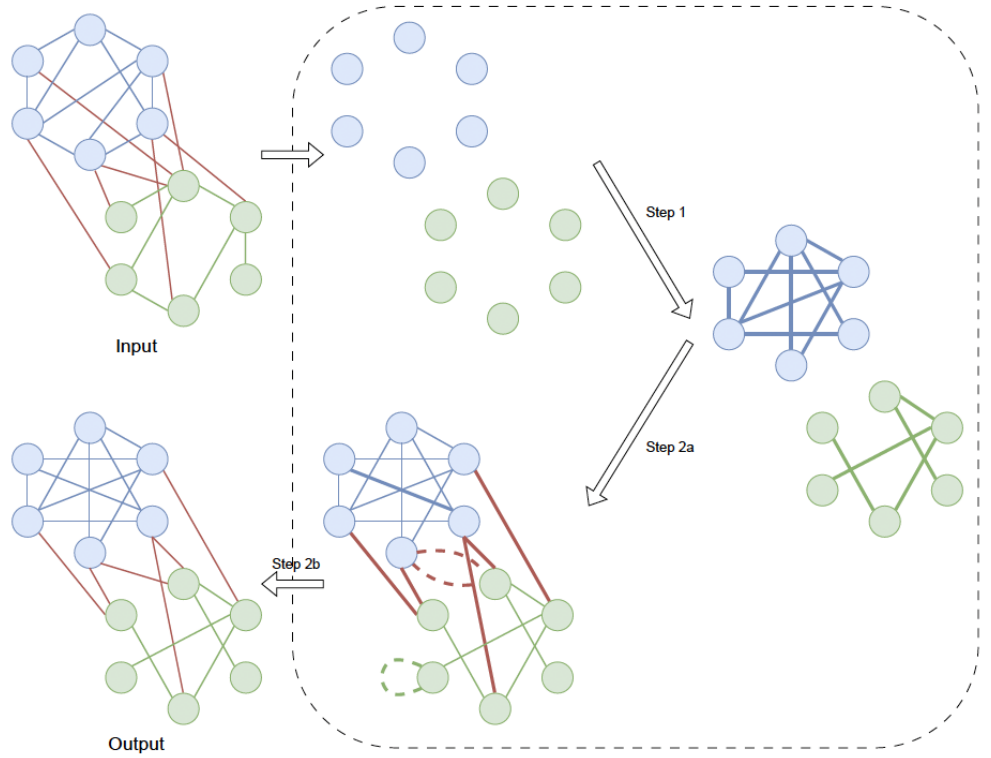


Fig. 2: **Stage 1 of EC-SBM: Generation of the synthetic clustered subnetwork**. The empirical cluster assignment is maintained as the synthetic cluster assignment. In Step 1, we generate for each cluster a k-edge-connected subnetwork where k is the desired edge connectivity of that cluster. In Step 2a, we generate the remaining edges according to the updated parameters using SBM; this can result in parallel edges and self-loops (dashed). In Step 2b, we remove the excessive edges to obtain the final output.

Table 1: Statistics used to evaluate synthetic networks

Statistic	Type and range of value(s)	Description
`pseudo_diameter` (*)	Scalar $(0, \infty)$	Approximate diameter
`char_time` (*)	Scalar $(0, \infty)$	Characteristic time of a random walk
`global_ccoeff` (*)	Scalar $[0, 1]$	Global clustering coefficient
`degree` (*)	Sequence $[0, \infty)$	Degree of the vertices
`mixing_mus`	Sequence $[0, 1]$	Local mixing parameter of each vertex
`mincuts`	Sequence $[0, \infty)$	Edge connectivity of each cluster
`c_edge`	Sequence $[0, \infty)$	Number of edges inside each cluster
`o_deg`	Sequence $[0, \infty)$	Degree of the outliers

These statistics are used to evaluate the fit between a synthetic network and a clustered real-world network. The first four statistics are network-specific and the last four are cluster-specific. The first four statistics are marked by (*), indicating that they do not depend on the cluster assignment.

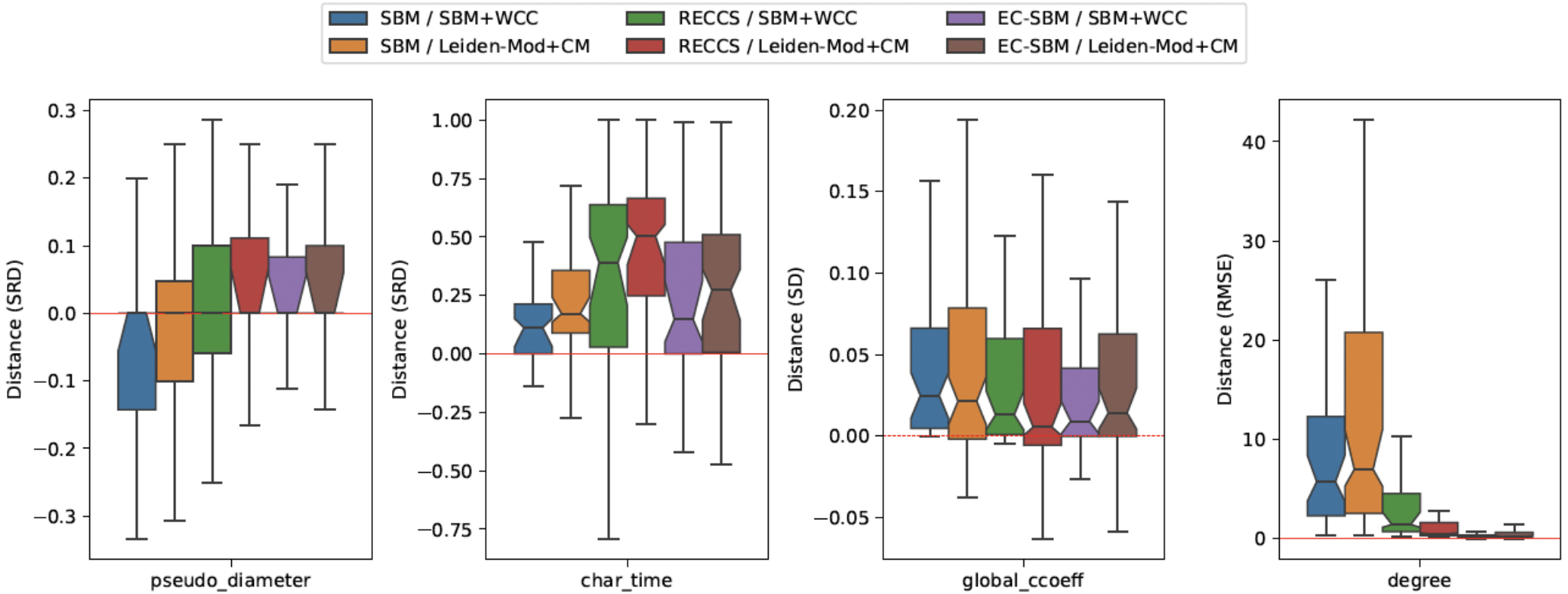


Fig. 6: Comparison between EC-SBM, SBM, and RECCS using two input clusterings on network-only criteria. The input clusterings are SBM+WCC and Leiden-Mod+CM, which we determine to be the most suitable in Experiment 1. The comparison is done on 74 networks with respect to 4 network-only criteria.

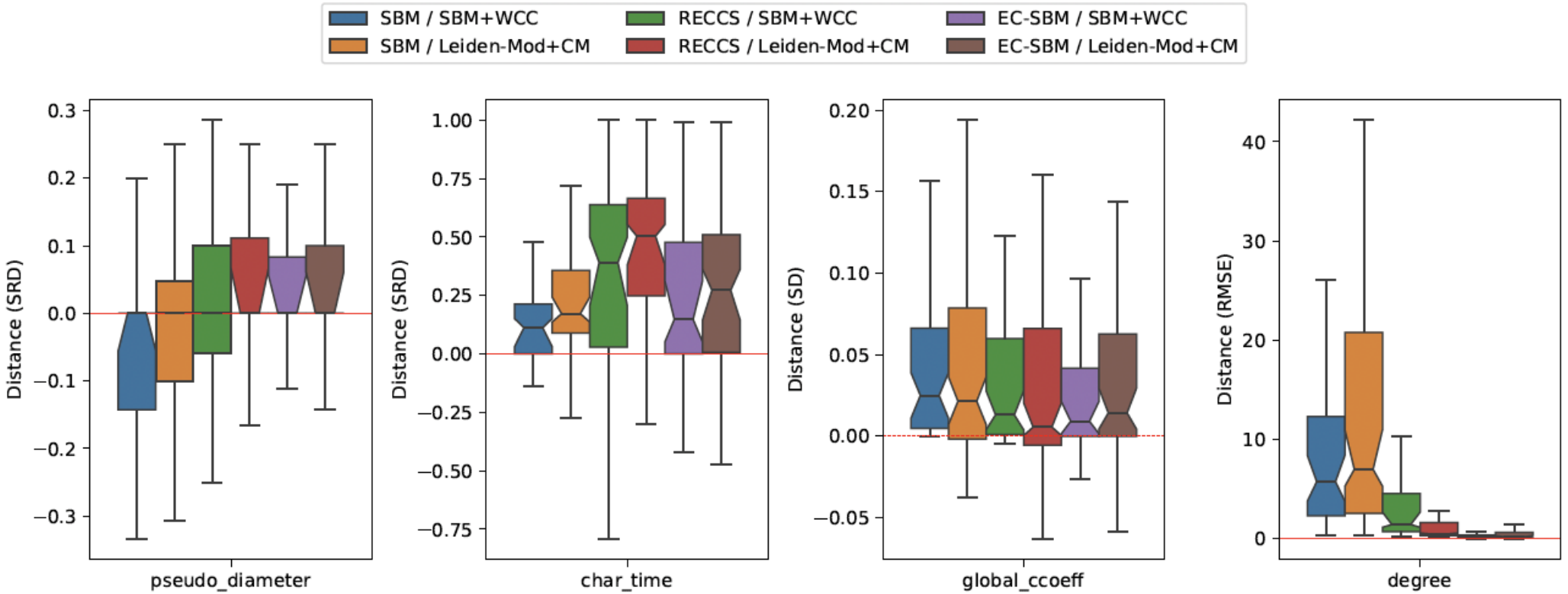


Fig. 6: Comparison between EC-SBM, SBM, and RECCS using two input clusterings on network-only criteria. The input clusterings are SBM+WCC and Leiden-Mod+CM, which we determine to be the most suitable in Experiment 1. The comparison is done on 74 networks with respect to 4 network-only criteria.

Note: Using SBM+WCC parameters produces the most accurate synthetic networks

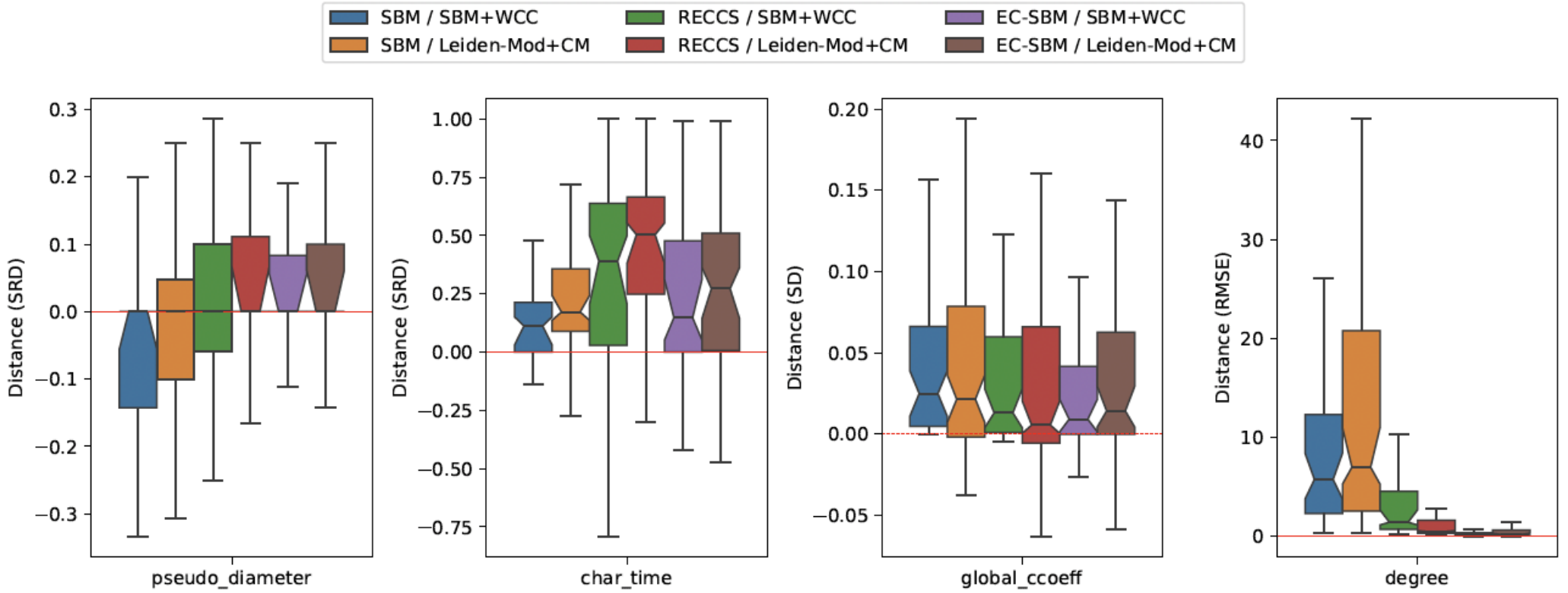
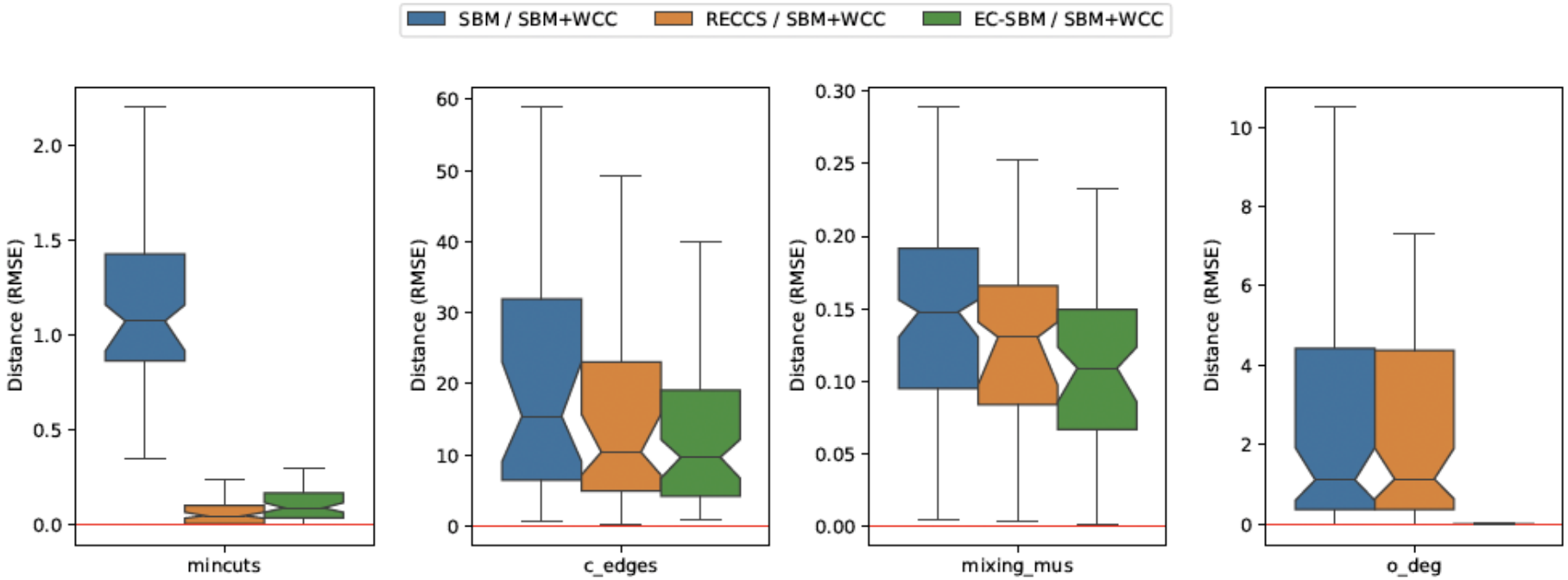


Fig. 6: Comparison between EC-SBM, SBM, and RECCS using two input clusterings on network-only criteria. The input clusterings are SBM+WCC and Leiden-Mod+CM, which we determine to be the most suitable in Experiment 1. The comparison is done on 74 networks with respect to 4 network-only criteria.

Note: EC-SBM and RECCS better than SBM on 3 out of 4 criteria

Comparison w.r.t. cluster properties



- RECCS and EC-SBM improve on DC-SBM for cluster properties
- EC-SBM better than RECCS for 3 of 4 criteria
- RECCS better than EC-SBM for mincuts

Table 2: Runtime analysis of the generation process of SBM, RECCS, and EC-SBM.

	CEN	orkut	livejournal
SBM (hours)	≈ 1.17	≈ 1.44	≈ 0.52
RECCS (hours)	≈ 1.62	≈ 3.50	≈ 2.55
EC-SBM (hours)	≈ 5.00	≈ 3.77	≈ 1.47
Number of nodes	14.0M	3.1M	4.9M
Number of outliers	12.8M	0.6M	2.0M
Number of edges	92.1M	117.2M	42.9M
Number of clusters	12K	31K	134K

Summary

- EC-SBM and RECCS are both more accurate than graph-tool SBM for cluster properties
- EC-SBM, RECCS, and SBM are similar for network properties
- SBM is fastest, but RECCS and EC-SBM are not much slower
- Choice between EC-SBM and RECCS depends on what properties are most important

Summary of trends

- Cluster connectivity is important for clustering and synthetic network generation
- Yet – standard methods fail to produce well-connected clusters
  - And some methods produce disconnected clusters!
- Graph-tool degree-corrected SBM favors a small number of clusters (a kind of “resolution limit”), leading it to produce disconnected clusters
- Simple ad hoc techniques are helpful in ameliorating these problems
  - Clustering: CM and WCC improve clustering accuracy (demonstrated on synthetic networks)
  - Synthetic networks: EC-SBM and RECCS both improve on SBMs

Conclusions

- Edge-connectivity is a natural expectation of valid communities.
- Off-the-shelf methods can produce disconnected or poorly-connected clusters.
- Clusterings should be examined.
- Some progress on improving cluster connectivity using simple methods (CM, WCC), and in producing better synthetic networks.
- Rigorous mathematical approaches and models are needed.

Software on github at https://github.com/illinois-or-research-analytics .

Papers available at https://tandy.cs.illinois.edu/bibliometrics.html .

Acknowledgments

- George Chacko
- Other collaborators: Fabio Ayres and Dmitriy Korobskiy
- Graduate students: Lahari Anne, Minhyuk Park, Vikram Ramavarapu, Baqiao Liu, Vidya Kamath Pailodi, Rajiv Ramachandran, and The-Anh Vu-Le
- Undergraduate students: Daniel Feng and Siya Digra
- Supported by the
  - National Science Foundation CISE/OAC 2402559,
  - Insper-Illinois Partnership,
  - Oracle, Digital Science, Google, and
  - The Grainger Foundation