Edge Connectivity and Stochastic Block Models (for Community Detection)

Tandy Warnow

Presentation for CS 598,

Feb 4 2025

Two papers

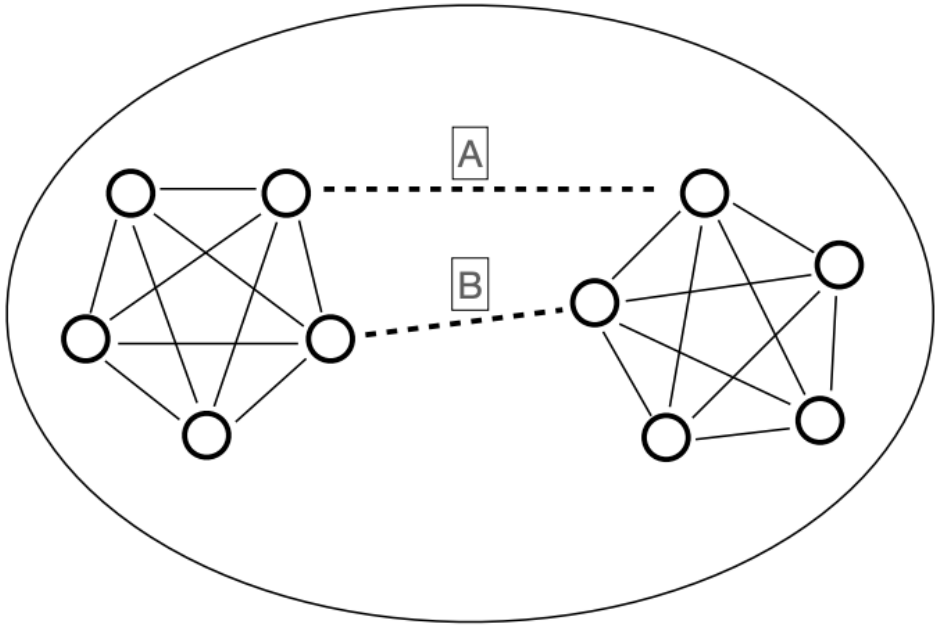
- Improved Community Detection using Stochastic Block Models. Minhyuk Park et al. Presented at Complex Networks and their Applications. Invited journal submission to PLOS Complex Systems
- Well-connectedness and community detection. Minhyuk Park et al. PLOS Complex Systems 2024;1(3):e0000009 (journal version of Complex Networks and their Applications 2023 paper)

Community Detection

- Given a network, partition the nodes into pairwise disjoint sets, so that each set has strong community structure:
  - Dense
  - Separated from the rest of the network
  - Well-connected (no small edge cut)

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

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

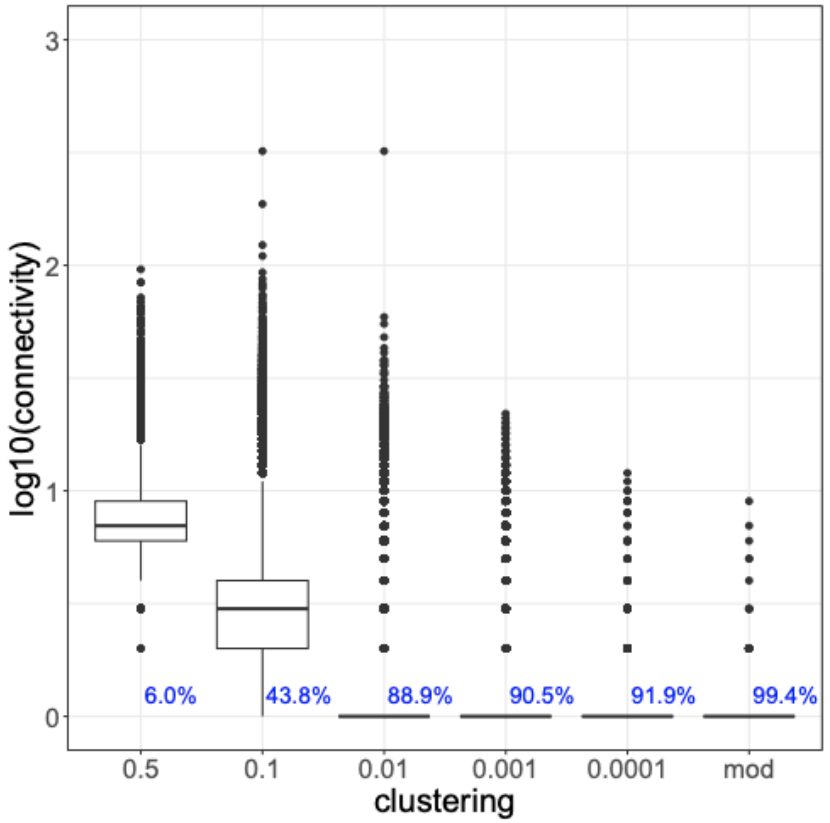
Theorem (rephrased from Traag et al. 2019):
Let C be a cluster in an optimal CPM clustering for resolution parameter γ
Suppose removing edge set E' splits C into sets X and Y.
Then E' has at least $\gamma |X||Y|$ edges.

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

The Connectivity Modifier

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

Leiden clusters have small edge cuts



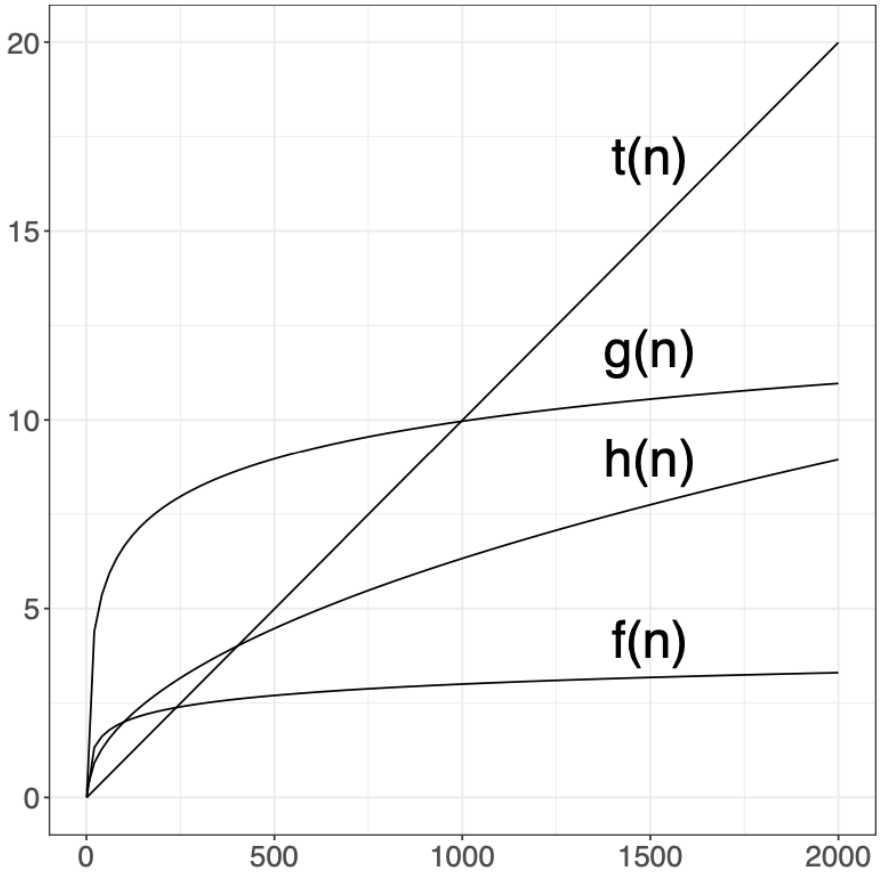
Leiden optimizing either Modularity (mod) or the Constant Potts Model (CPM) for different resolution values.

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).*

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



n = cluster size

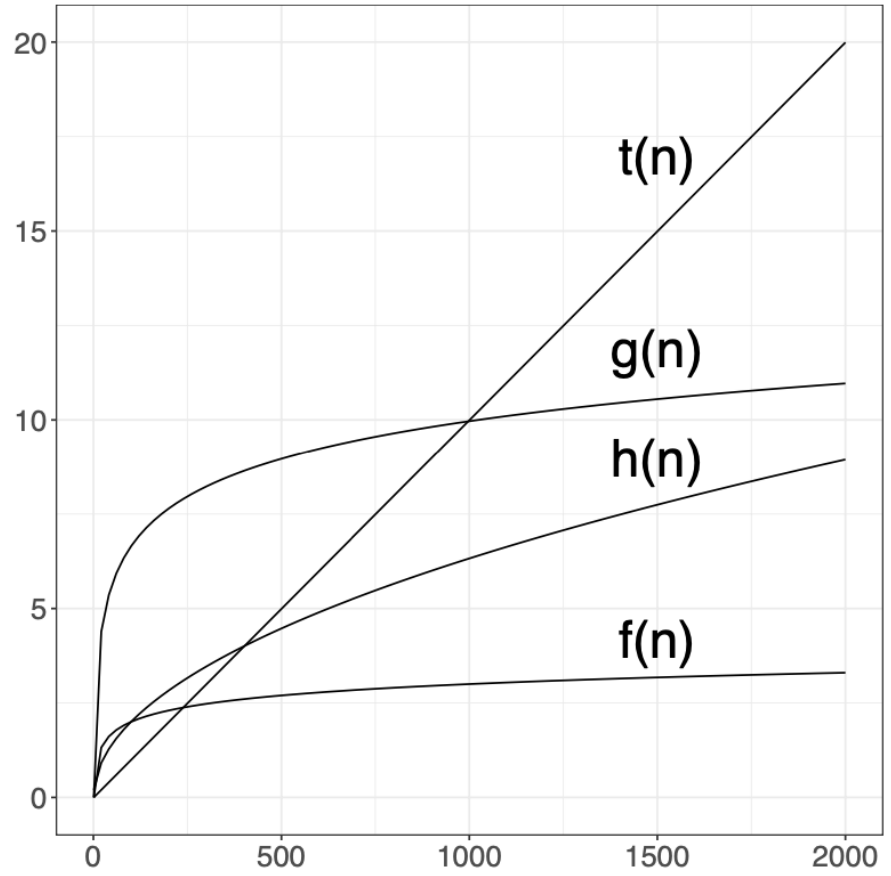
$f(n) = \log_{10} n$

$g(n) = \log_2 n$

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

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

We use f(n) for “well-connected”

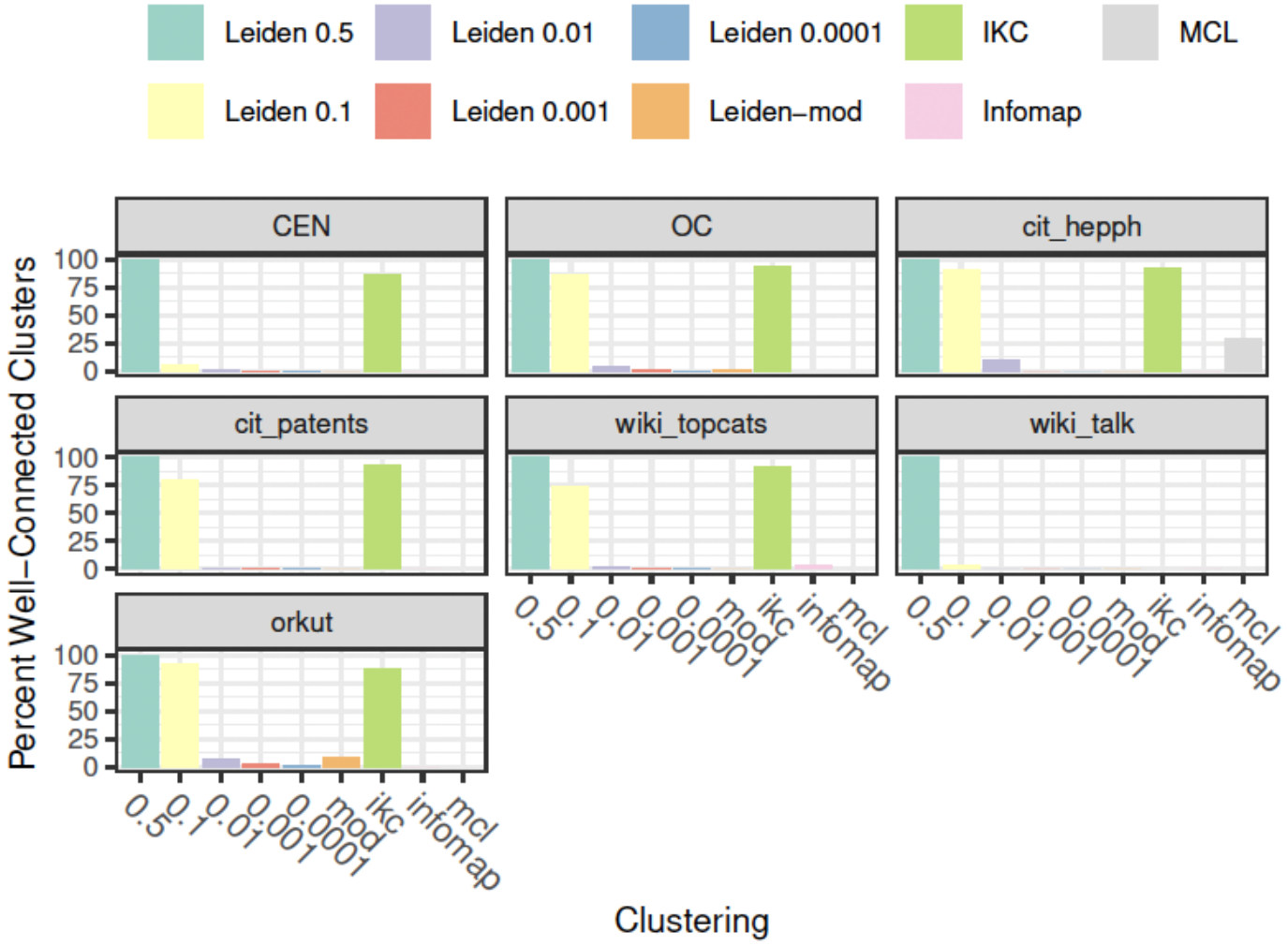


n = cluster size

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

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

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

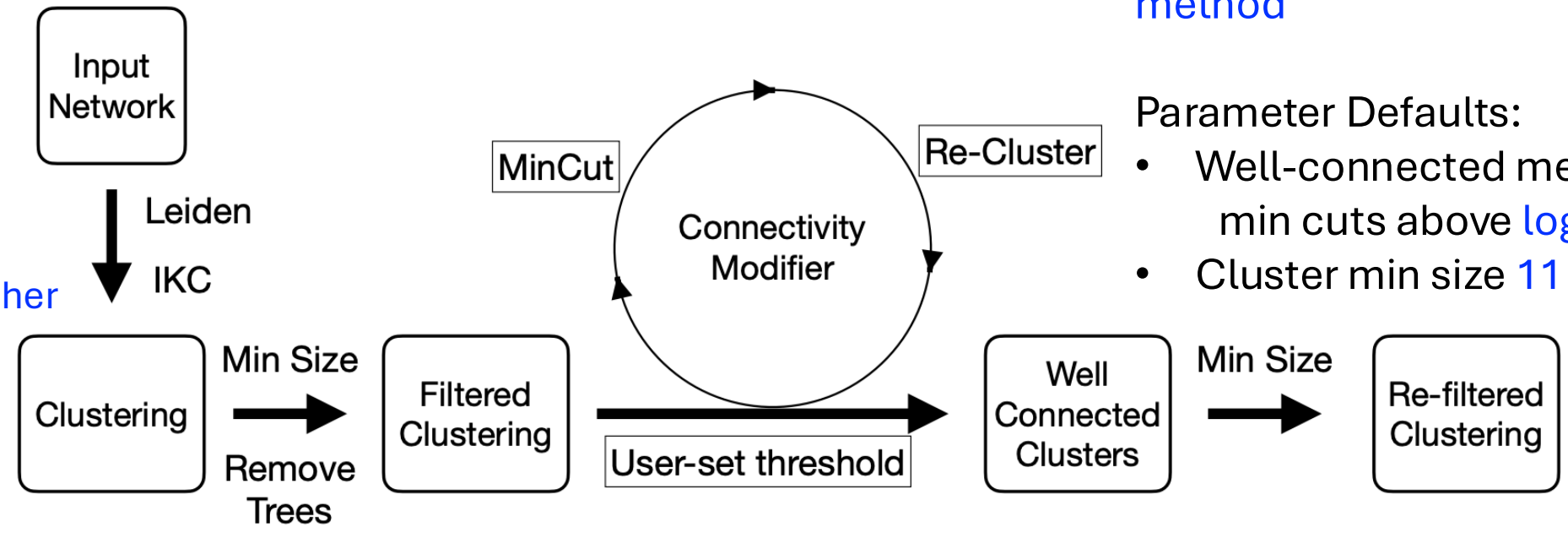
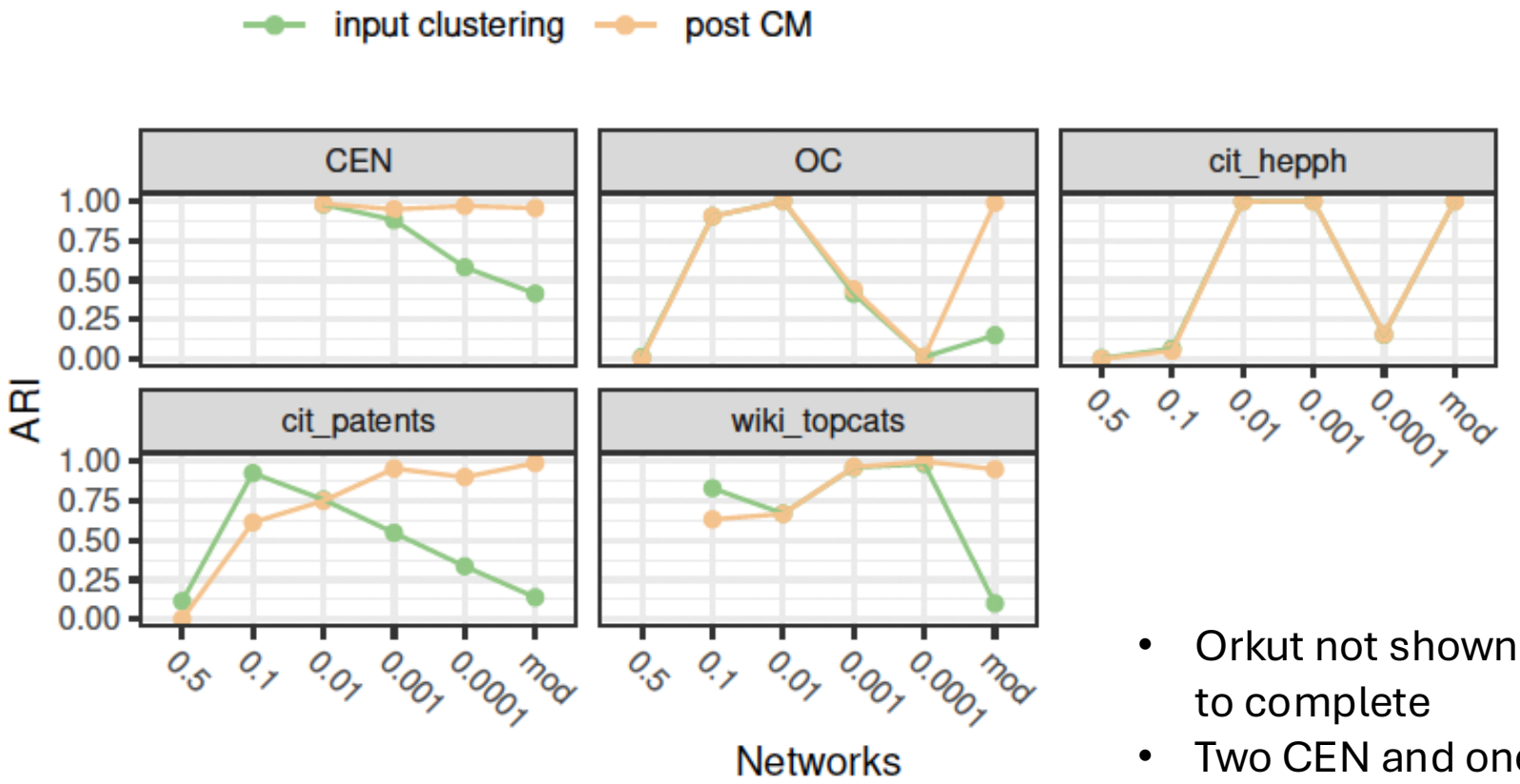


Figure 3: *Connectivity Modifier Pipeline Schematic* The four-stage pipeline depends on user-

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

After my talk at CNA 2023, I was asked:
What about Stochastic Block Models??

- Popular generative model for networks with community structure
- Can be used to generate synthetic networks or to find community structure in a network
- SBMs in wide use using graph-tool by Peixoto

- The question I was asked was specifically asking: Do SBMs also have poorly connected clusters, and if so, could the Connectivity Modifier help?

Improved Community Detection using Stochastic Block Models

Park et al., Complex Networks and their Applications 2024

(submitted, by invitation, to PLOS Complex Systems)

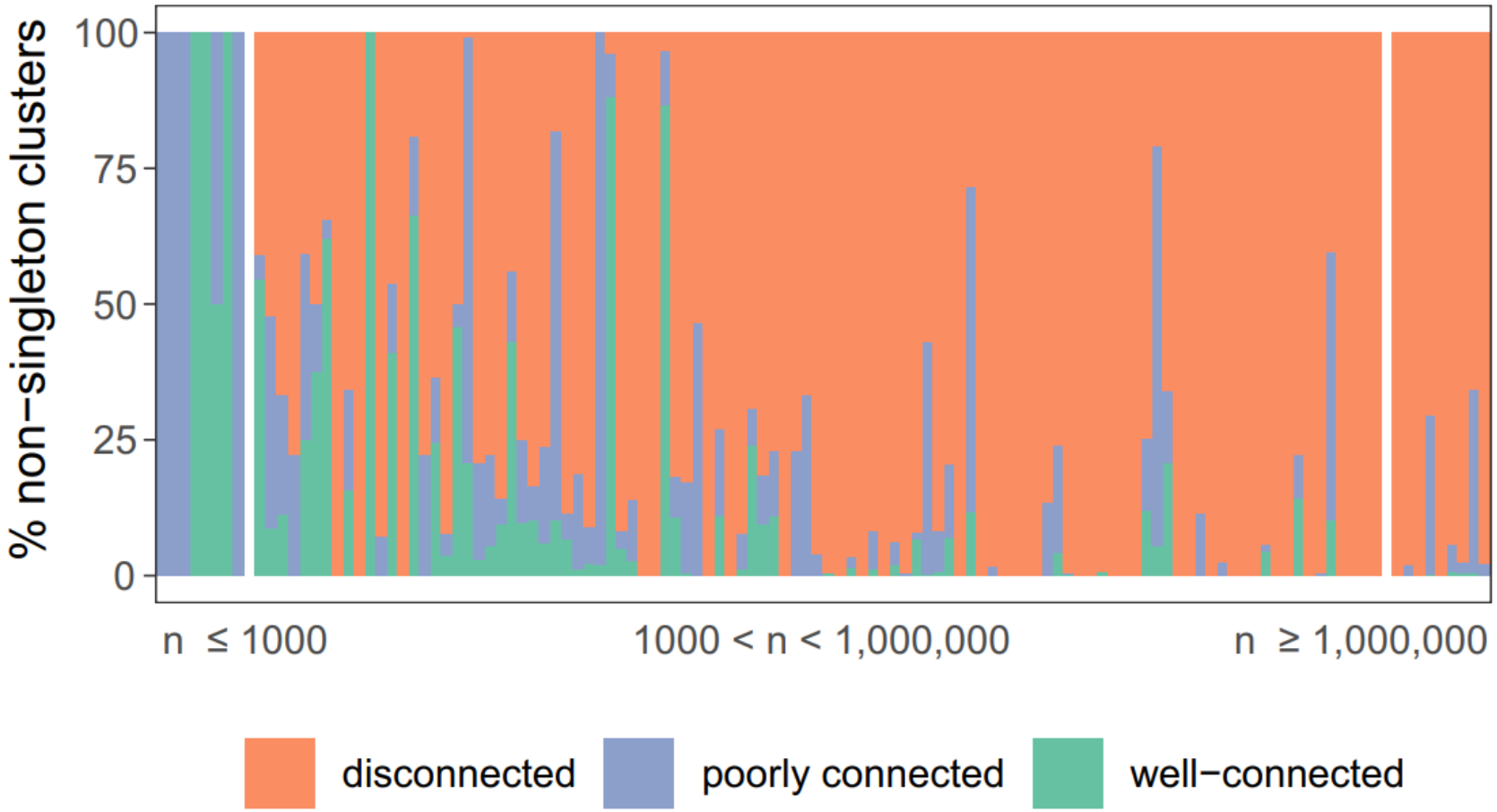
- Evaluated clustering using SBM for connectivity on 120 real-world networks
- Examined impact on accuracy on LFR networks (from CM paper) using three treatments
  - CM (Connectivity modifier) – but without filtering for size
  - Well-Connected Clusters (WCC) and
  - Connected Components (CC)

Different models of SBM

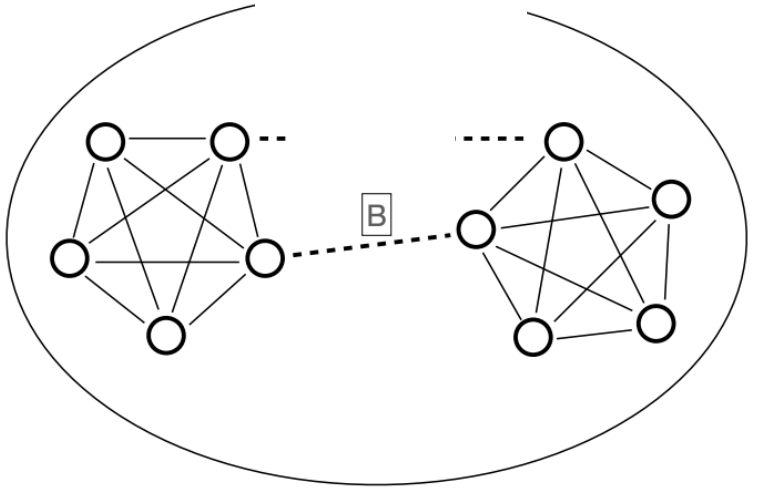
- Several SBM models are available in the graph-tool package (Peixoto):
  - Degree-corrected
  - Non degree-corrected
  - Planted partition

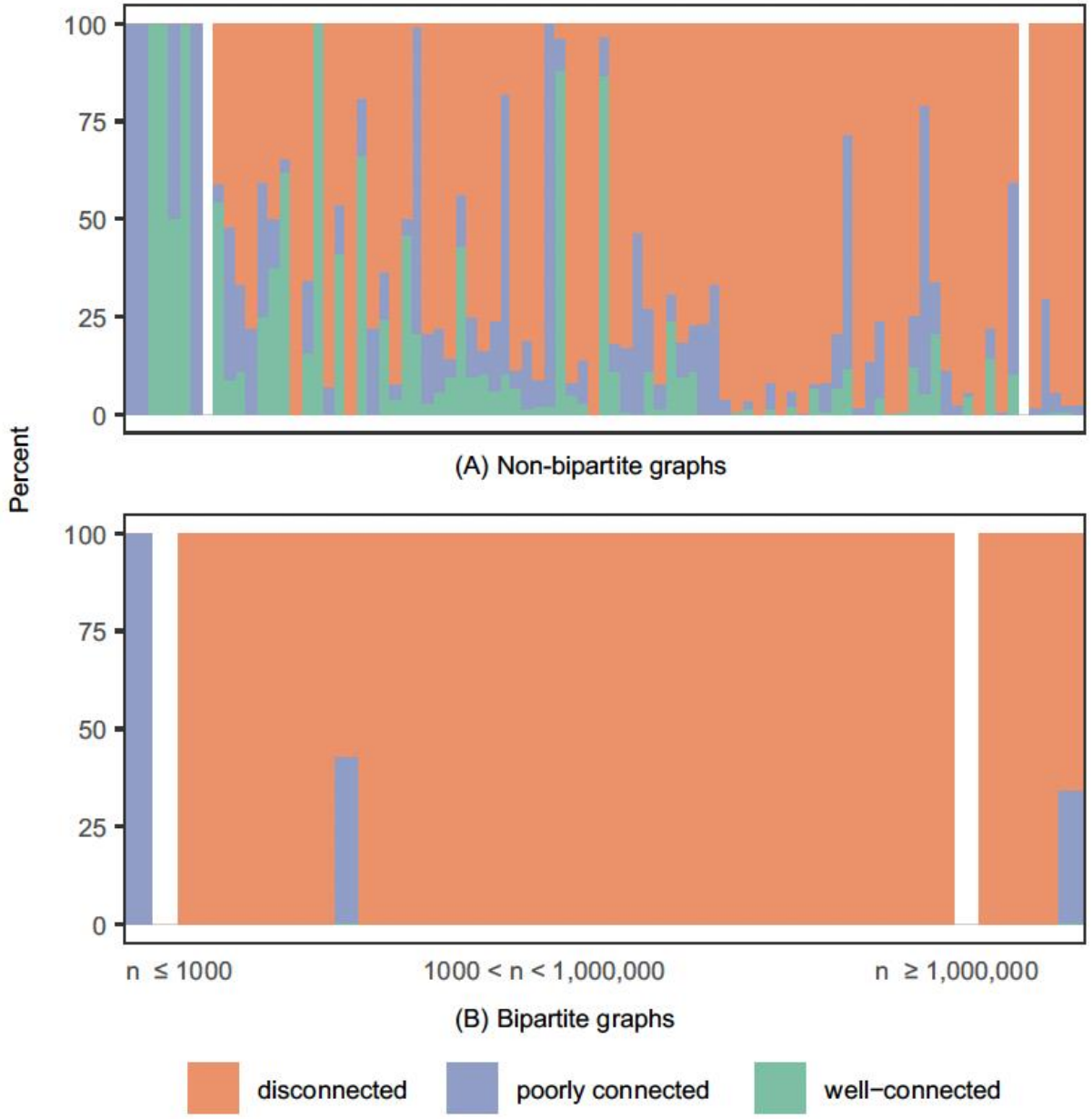
- Protocol:
  - Cluster an input network using all three models
  - Compute the description length (fitness of clustering to input data) for all three
  - Choose the clustering with the minimum description length

SBM clustering of real-world networks



- Stochastic Block Model clusterings often produce **disconnected** clusters
- Results shown are on 120 real world networks

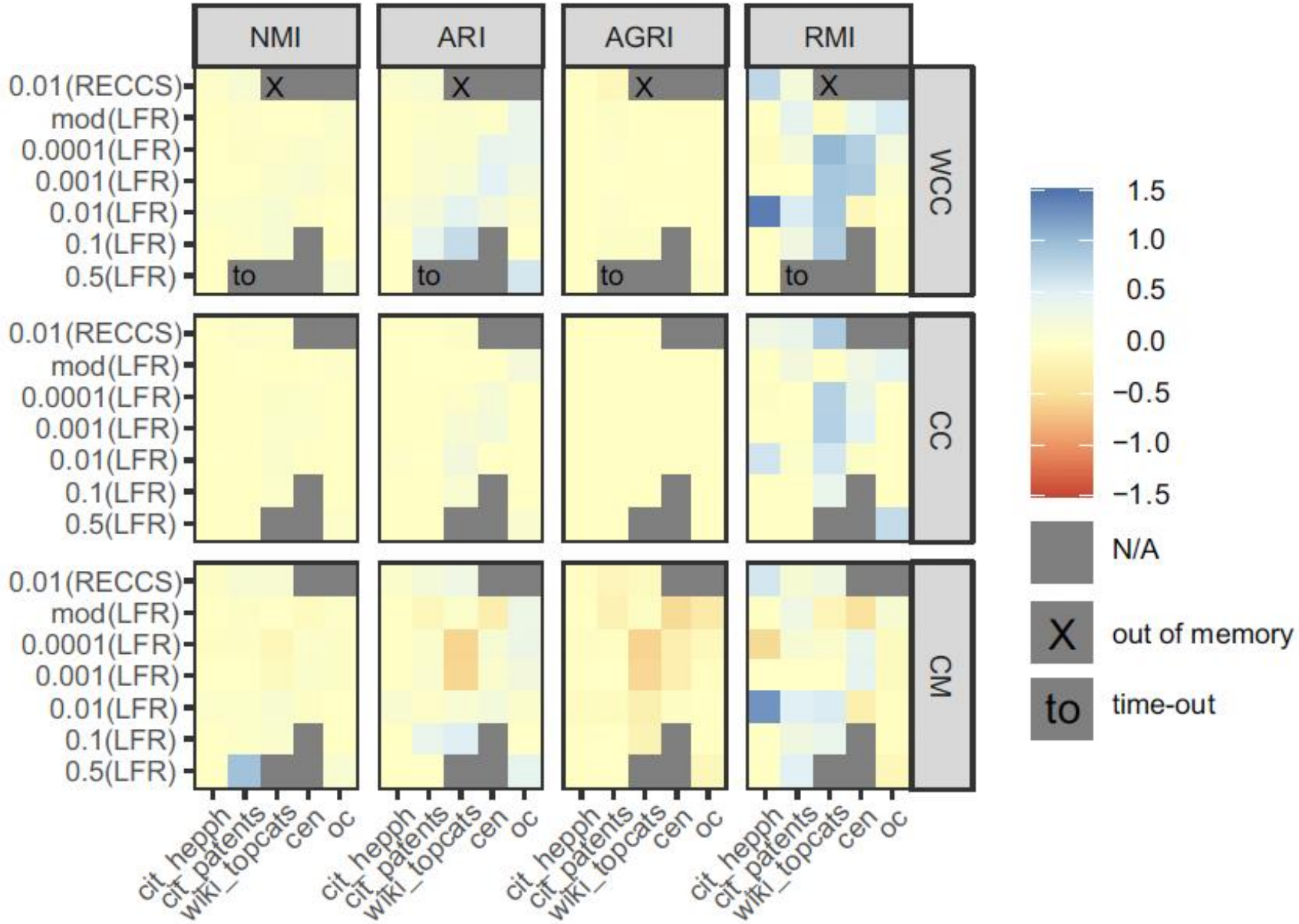




Much worse for bipartite graphs

Three techniques to improve edge connectivity

- Connected Components (CC): return connected components
- Well Connected Clusters (WCC):
  - If a cluster is not well-connected
    - Remove small edge cut to split into two parts
    - Recurse
- Connectivity Modifier (CM): repeatedly
  - If a cluster is not well-connected
    - Remove small edge cut to split into two parts
    - Recluster each part
    - Recurse



Four accuracy measures

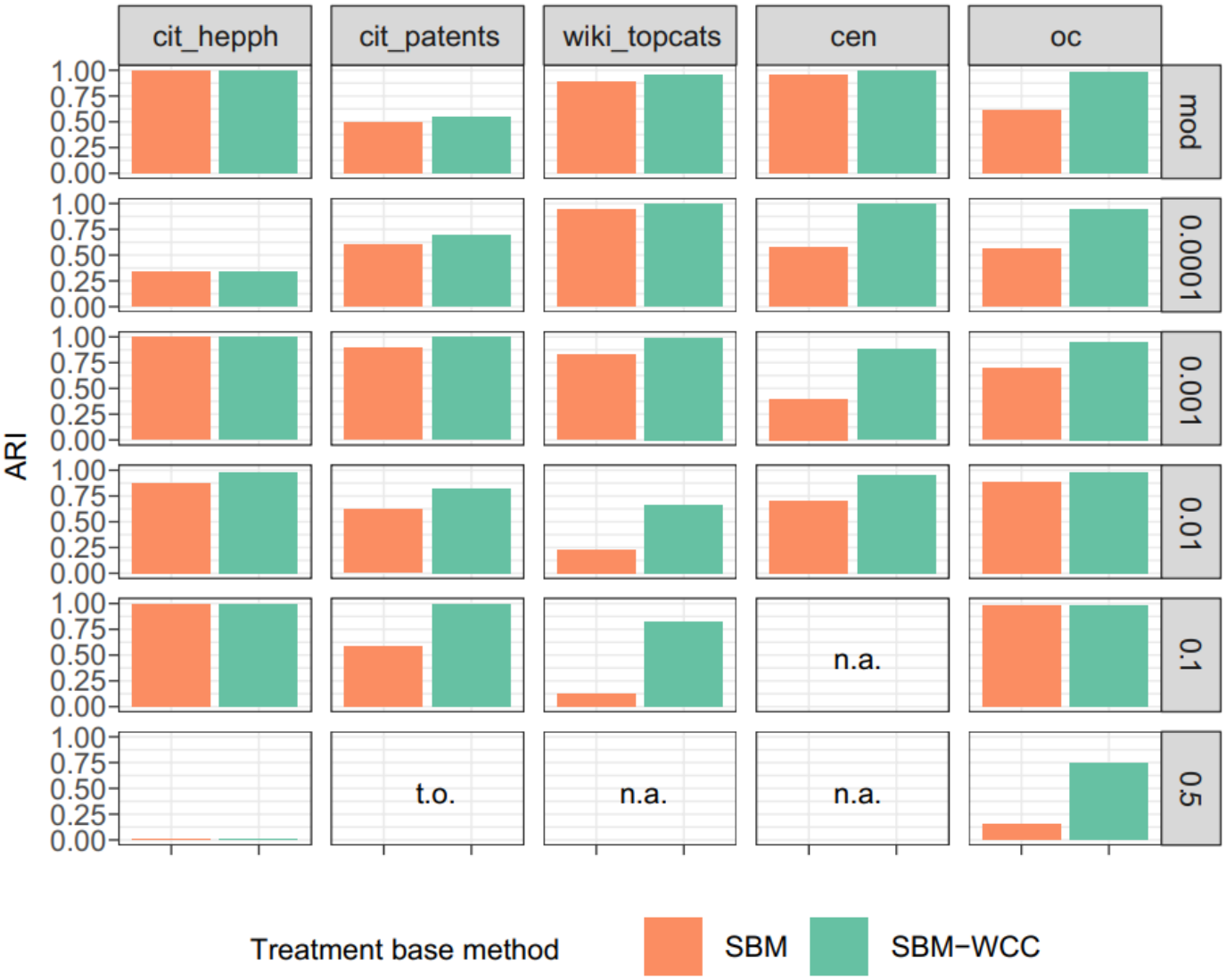
35 synthetic networks

Compared SBM+WCC, SBM+CC, and SBM+CM, and untreated SBM

Color indicates if treatment helped, hurt, or was neutral

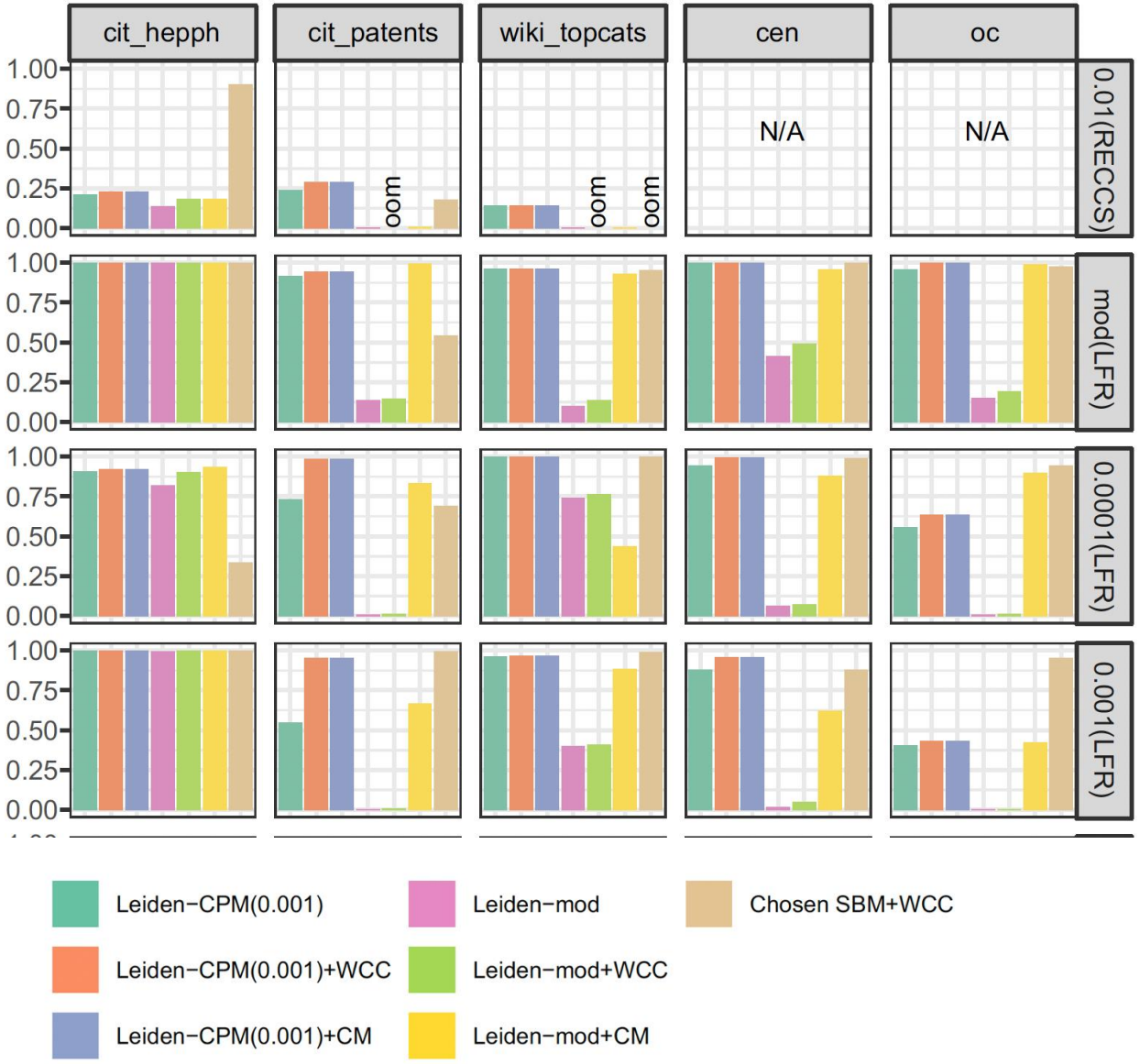
WCC is the best of these treatments!

Impact of WCC on SBM accuracy on LFR



- WCC treatment improves SBM accuracies
- Small improvements tend to be those with already high accuracy
- Same LFR networks as CM study (CNA 2023)

Comparison of SBM+WCC to Leiden-Mod and Leiden-CPM



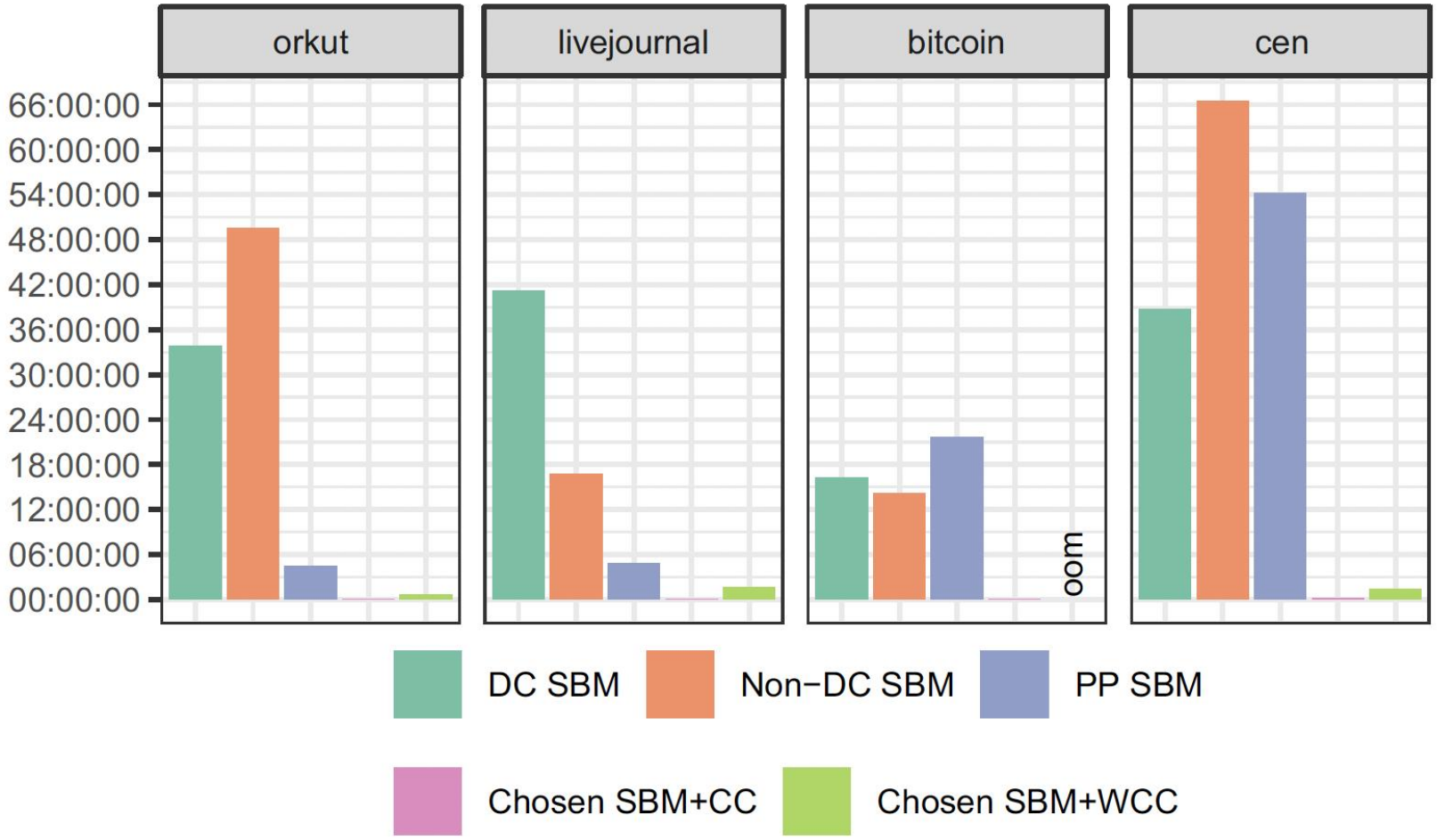
NMI accuracy

Leiden optimizing CPM or Modularity

20 synthetic networks

SBM+WCC competitive with the best!

Runtime


orkut
livejournal
bitcoin
cen
66:00:00
60:00:00
54:00:00
48:00:00
42:00:00
36:00:00
30:00:00
24:00:00
18:00:00
12:00:00
06:00:00
00:00:00
oom
DC SBM
Non−DC SBM
PP SBM
Chosen SBM+CC
Chosen SBM+WCC

Why does SBM clustering produce disconnected clusters?

- SBM+CC is neutral to beneficial, and SBM+WCC is generally beneficial for clustering accuracy.
- But SBM clustering produces disconnected clusters
- Why?
- The key is understanding the optimization problem, minimizing the description length.

The description length formula

- 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:

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

Table 1. Components of description lengths on `linux`.

Quantity	SBM(DC)	SBM(DC)+CC	Difference
$-\log p(A\|b,e,k)$	699k	316k	-383k
$-\log p(k\|b,e)$	96k	45k	-51k
$-\log p(b)$	147k	257k	110k
$-\log p(e)$	51k	1,585k	1,534k
$\mathrm{DL}(A,b)$	993k	2,202k	1,209k

The last row sums up the values in the previous four rows. The difference is SBM(DC)+CC - SBM(DC); so, a negative value means favoring SBM(DC) with CC treatment and a positive value means favoring untreated SBM(DC).

Trends across many networks:
P(e) favors disconnected clusters

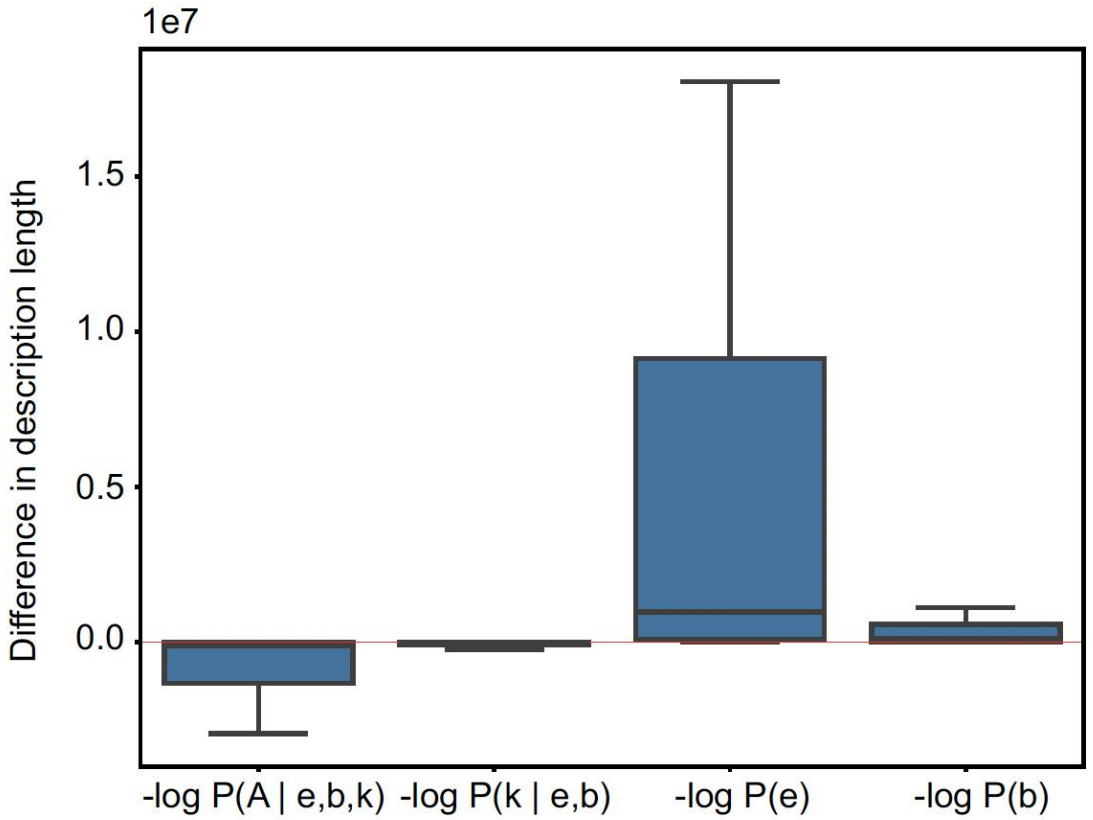


Fig 8. SBM(DC)+CC to SBM(DC) difference for description length

The description length formula

- 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:

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

Analyzing the log p(e) term

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

Easy to see that this grows with the number B of blocks, since the number E of edges is fixed

Effect of log p(e) term

We investigated all networks to see if the CC treatment is preferred when we remove the $-\log p(e)$ component. Our investigation shows, for 59 networks out of 71 networks, removing the $-\log p(e)$ component will result in a lower description length for the clustering output with CC treatment. Thus, the $-\log p(e)$ component accounts for 83.1% of the cases where SBM(DC) without CC treatment is preferred over SBM(DC) with CC treatment.

Take home points

- All tested clustering methods produced poorly-connected clusters.
- SBMs are even among the worst!
- The cause for SBMs may be the description length formula
- But why is this true for other clustering methods?
- 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