Improved Community Detection using Stochastic Block Models

Presenter: Minhyuk Park

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

University of Illinois Urbana-Champaign

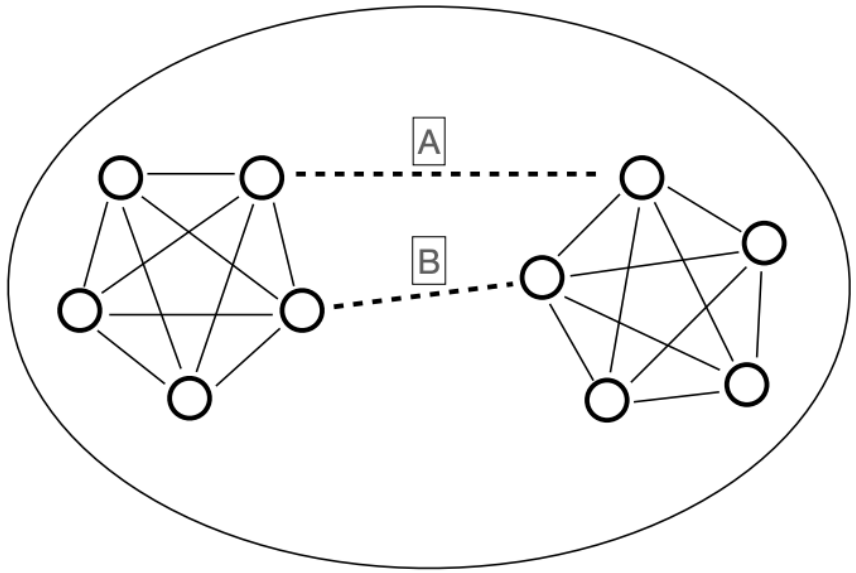
1 2 / 1 1 / 2 4

Funded by Illinois-Insper Partnership

Overview

- Background material
- Park et al. CNA 2023, Park et al. PLOS Complex systems 2024
- Improving SBM connectivity

Background - Edge Cut Example



- Edge cut:
  - Set of edges whose removal splits a graph into two components
  - Mincut is an edge cut with the smallest size
- Consider the cluster on the left:
  - No edge cuts of size 1
  - Edge cut of size 2: {A, B}
  - Mincut size is 2

Background – desirability of well-connectedness

- A large mincut is desirable (Kannan et al., "On clusterings: Good, bad and spectral." JACM 2004; Zhu et al., “A local algorithm for finding well-connected clusters.” ICML 2013)


SCIENTIFIC REPORTS

OPEN

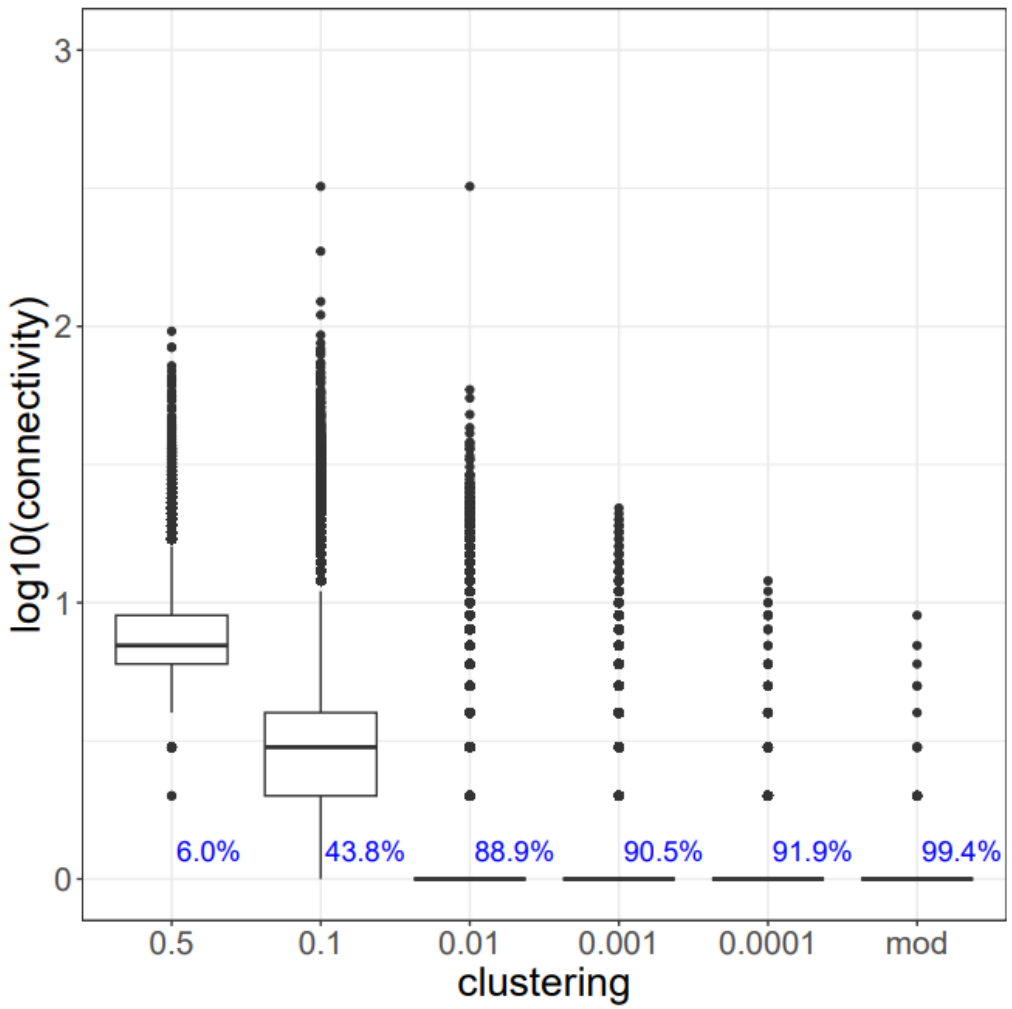
From Louvain to Leiden: guaranteeing well-connected communities

ed: 20 November 2018

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

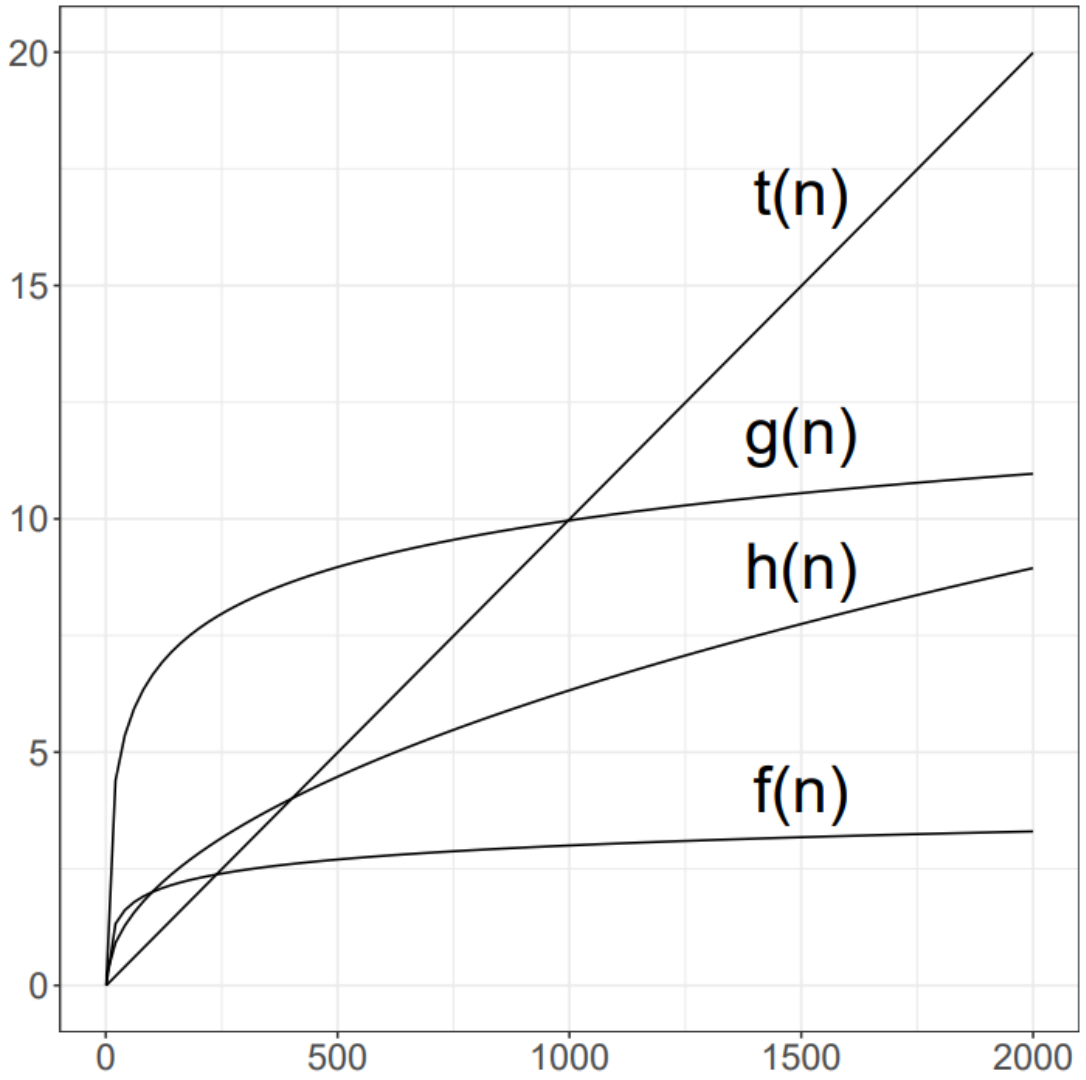
- Traag et al. proved that CPM-optimal clusterings satisfy the following:
  - if E is an edgecut splitting cluster into A and B and γ is the resolution parameter, then
  - $|E| \geq \gamma |A||B|$

Leiden-CPM Has Small Mincuts



- Results shown on Open Citations network
  - 75 million nodes
  - 1.4 billion edges
- Leiden clusterings on the x-axis:
  - Numeric = Leiden-CPM γ
  - Mod = Leiden-Mod
- Mincut sizes shown on the y-axis
- Blue text shows percentage of nodes in non-singleton clusters out of total nodes

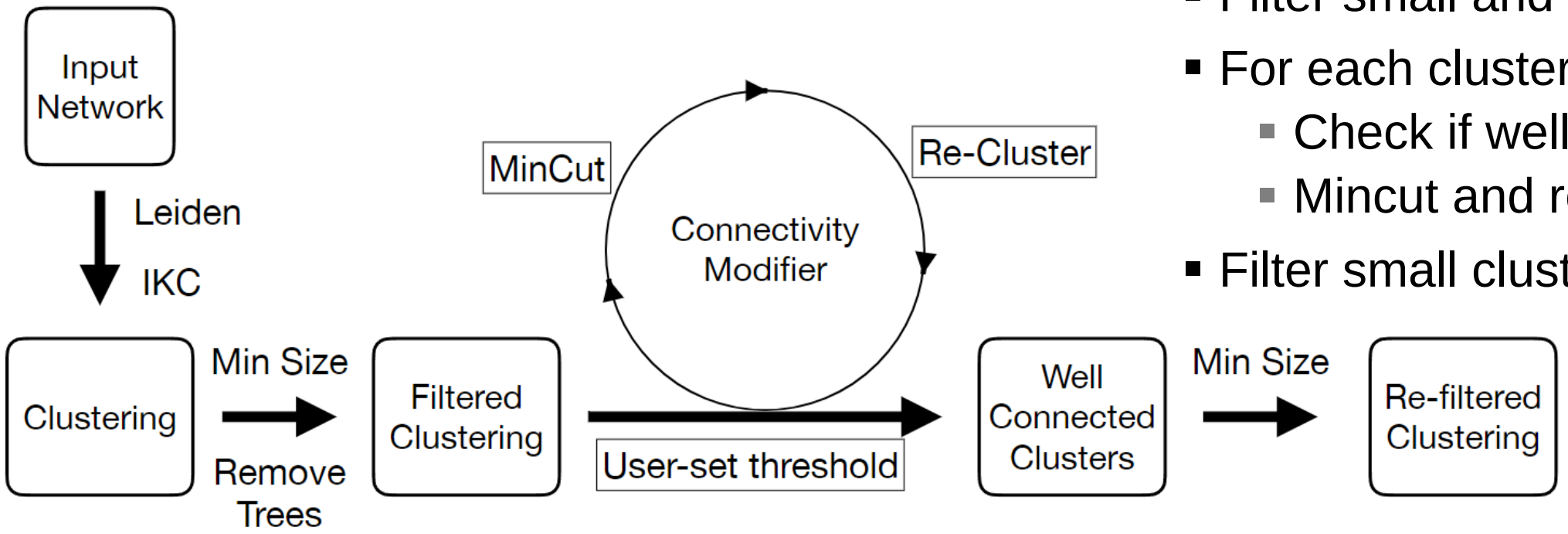
Choice of f(n)



- t(n):
  - Result from Traag et al. with γ=0.01
  - 0.01(n-1)
- f(n) = $\log_{10} n$
- g(n) = $\log_2$ n
- h(n) =$\frac{\sqrt{n}}{5}$

- f(n) larger than t(n) for small n
- f(n) smaller than t(n) for large n

Background – Connectivity Modifier



- Start with input clustering
- Filter small and tree clusters
- For each cluster:
  - Check if well-connected f(n)
  - Mincut and re-cluster
- Filter small clusters

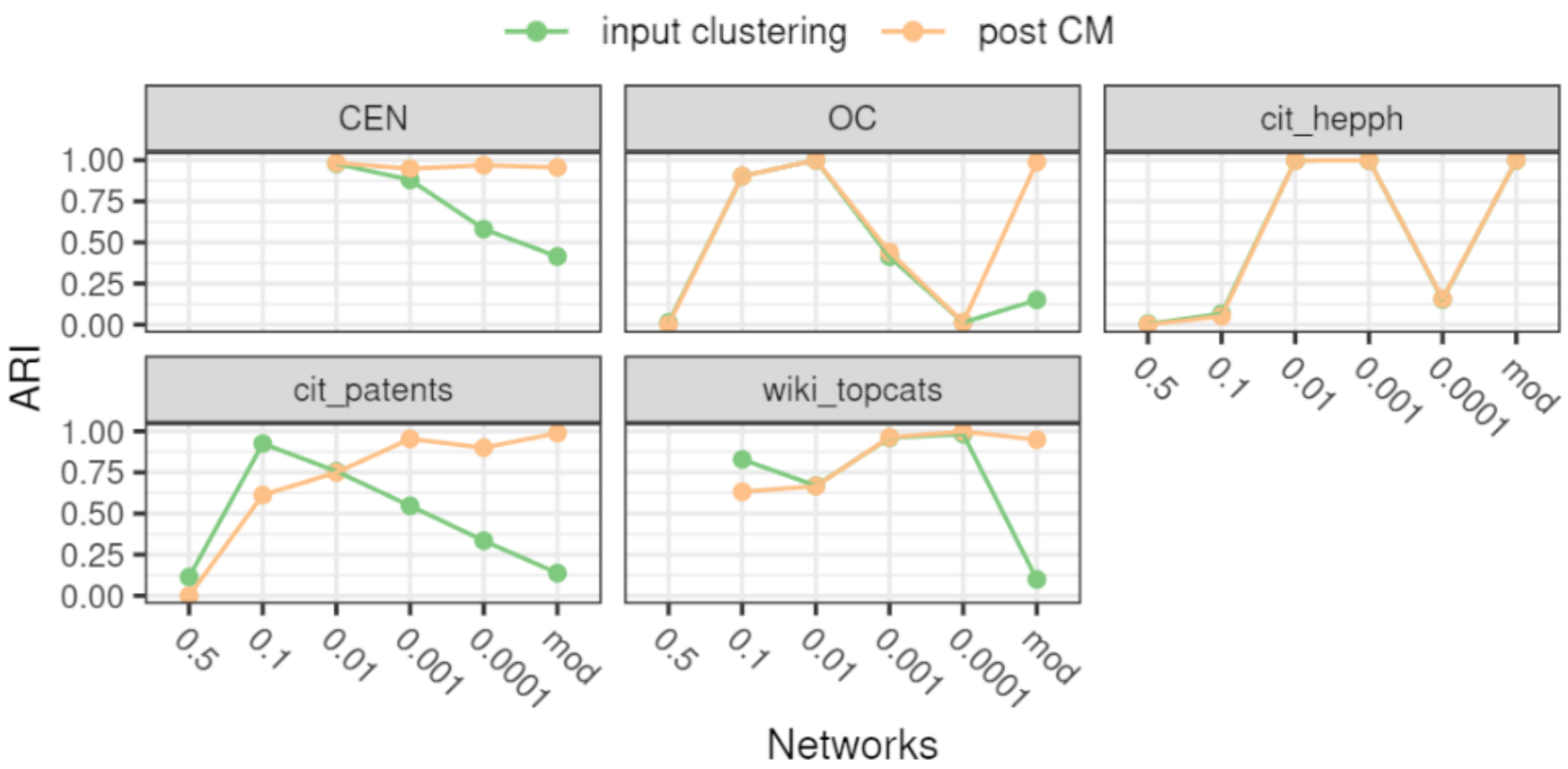
See also Ramavarapu et al. JOSS 2024

LFR Networks from Park et al. CNA 2023, PLOS Complex Systems 2024

- Network generation:
  - Compute numeric parameters based on an empirical network and clustering
  - Provide numeric parameters to LFR
  - Note: some LFR created networks were omitted
    - LFR failed to compute on CEN 0.1, 0.5 with provided parameters
    - wiki_topcats 0.5 and all wiki_talk -> disconnected ground truth clusters

- Experiments (evaluating impact of CM):
  - Re-cluster LFR network using the same clustering method
  - CM-processing with the same clustering method

Prior Literature - CM Insights



- CM processing can improve clustering accuracy
- Achieves this by splitting clusters to increase cluster connectivity

Motivation

- What about SBM-based clusterings?

- Our new study addresses the following:
  - Does SBM produce poorly connected clusters?
  - If so, can CM improve it?

Datasets

- Evaluation of SBM clusterings on 120 real-world networks:
  - Netzschleuder network catalogue and repository by Peixoto + 2 more
  - Network sizes range from 11 nodes to about 14 million nodes

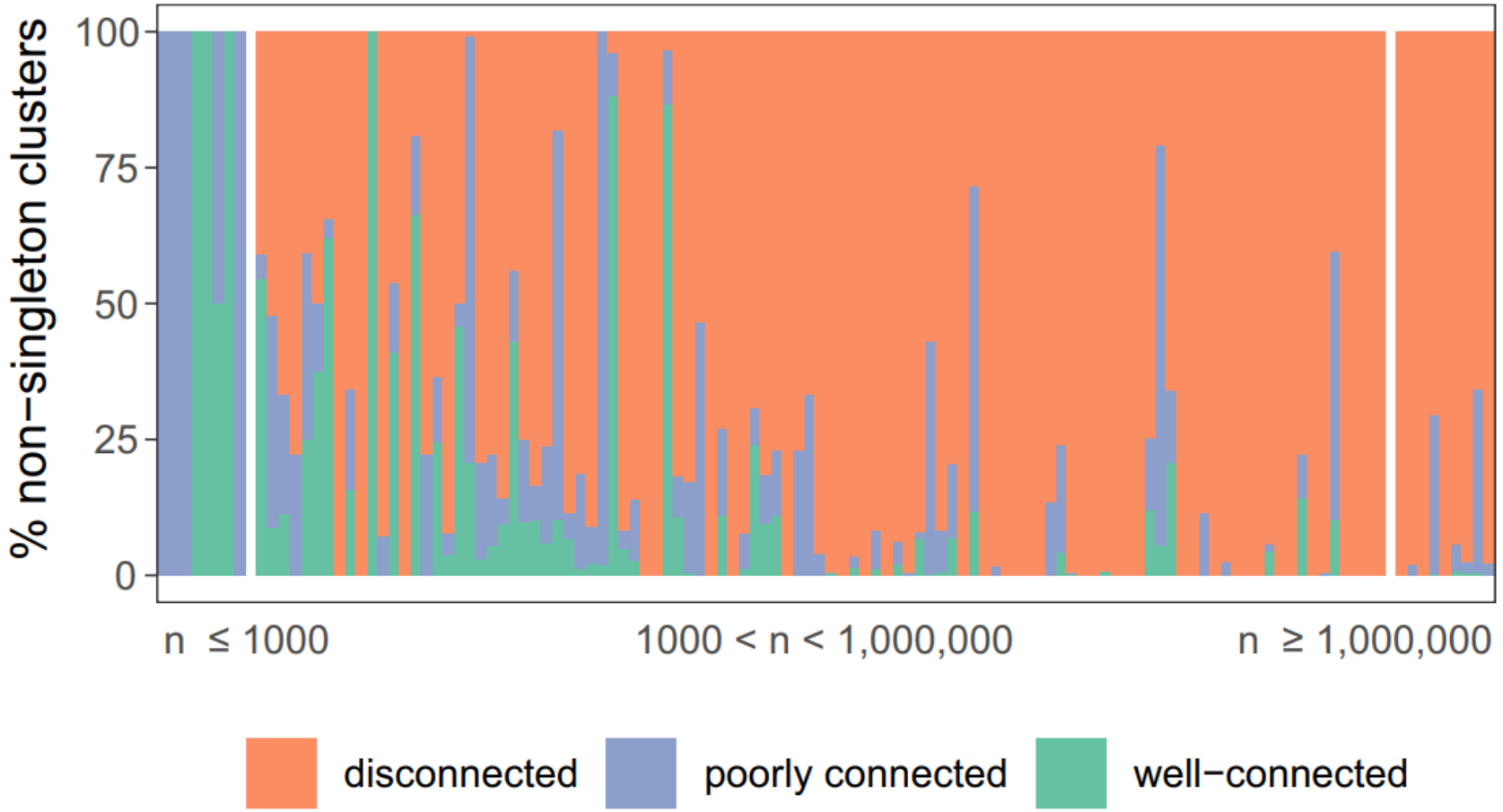
- Evaluation on LFR networks from Park et al. CNA 2023 (sizes up to ~3 million)

Different models of SBM

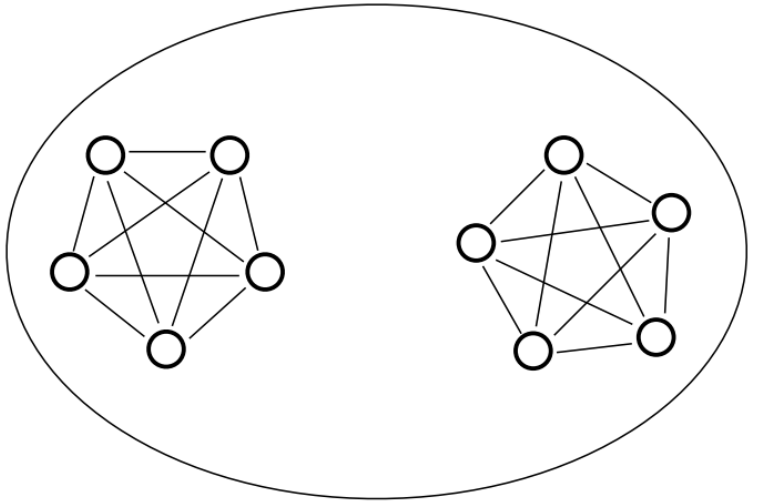
- Several SBM models are available in the graph-tool package (Peixoto):
  - Degree-corrected
  - Non degree-corrected
  - Planted partition

- 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

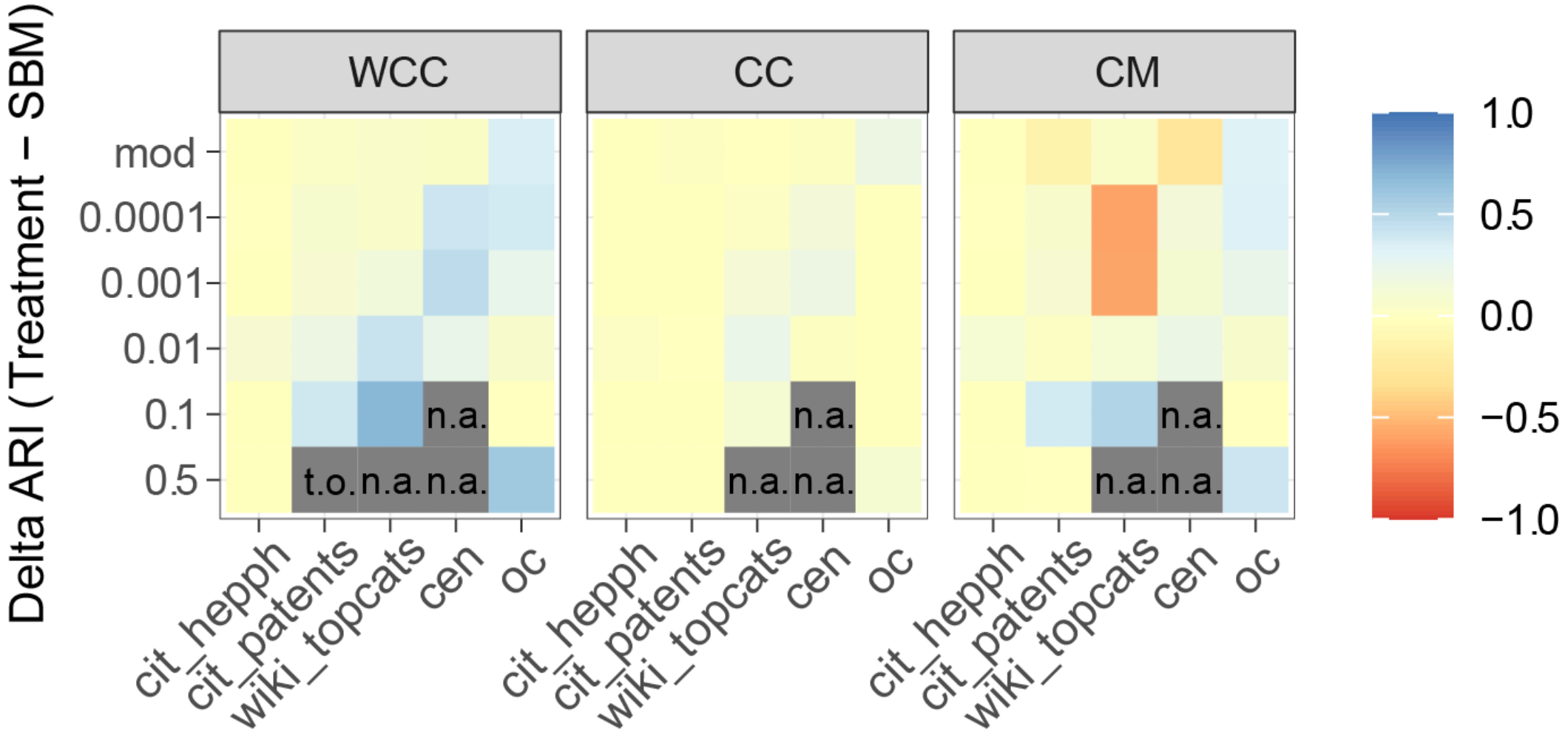


Post-clustering Treatments

- **CM** – Connectivity Modifier: Omitting filtering step
- **CC** – Connected Components: Return connected components of each cluster
- **WCC** – Well Connected Clusters: Repeated mincuts until all clusters are well-connected

- Evaluation treatment impact on NMI, ARI, AMI

Impact of treatment of SBM accuracy on LFR networks



Impact of WCC on SBM accuracy on LFR networks

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

Description Length of SBM

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

Description Length of SBM

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

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

- B = # blocks (clusters), E = # edges
- Increasing B produces large positive value - worse description length

Description Length of SBM

Quantity	SBM(DC)	SBM(DC)-CC
$-\log p(A\|b,e,k)$	699,228	315,645
$-\log p(k\|b,e)$	95,737	45,066
$-\log p(b)$	147,019	256,817
$-\log p(e)$	**50,786**	**1,584,555**
$\mathrm{DL}(A,b)$	992,771	2,202,083

- Description length penalizes having many clusters
- CC clusterings have worse description length
- $-\log p(e)$ is the reason for CC having worse description length on 80 out of 103 networks tested (77.7%)

Summary

- Clustering using SBM often produces disconnected clusters:
  - Minimum description length penalizes having many clusters

- WCC improves accuracy on synthetic networks but CM has variable impact

Future work

- More rigorous mathematical models
- Evaluation based on FNR, FPR, and AGRI (Poulin, V. and Théberge, F., IEEE Transactions on Pattern Analysis and Machine Intelligence 2020.)

Grainger College of Engineering

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN