Well-connected clusters: a valuable but elusive property in community detection methods

T. Warnow

University of Illinois Urbana-Champaign

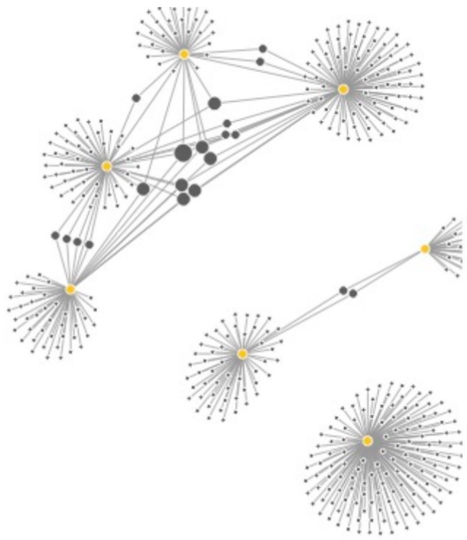
Joint work with George Chacko, Minhyuk Park, and others at UIUC, NTT Data, and Insper (Brazil)

Supported by: Illinois-Insper Partnership, Oracle Research Award, 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.



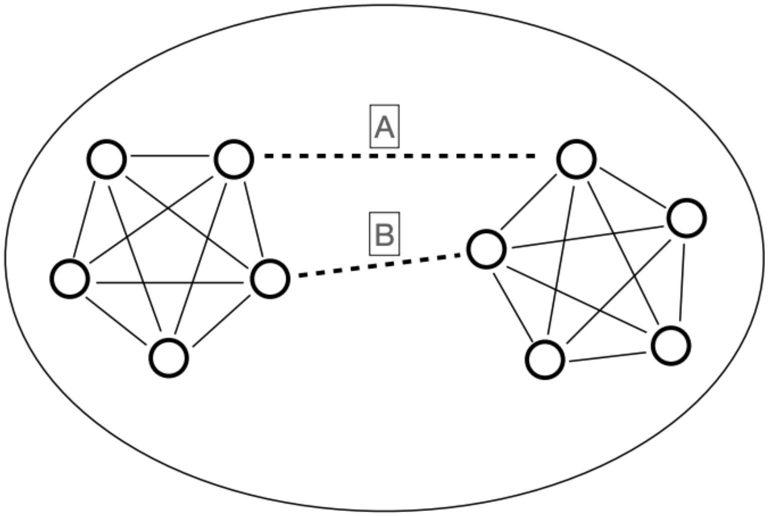
https://tandy.cs.illinois.edu/bibliometrics.html

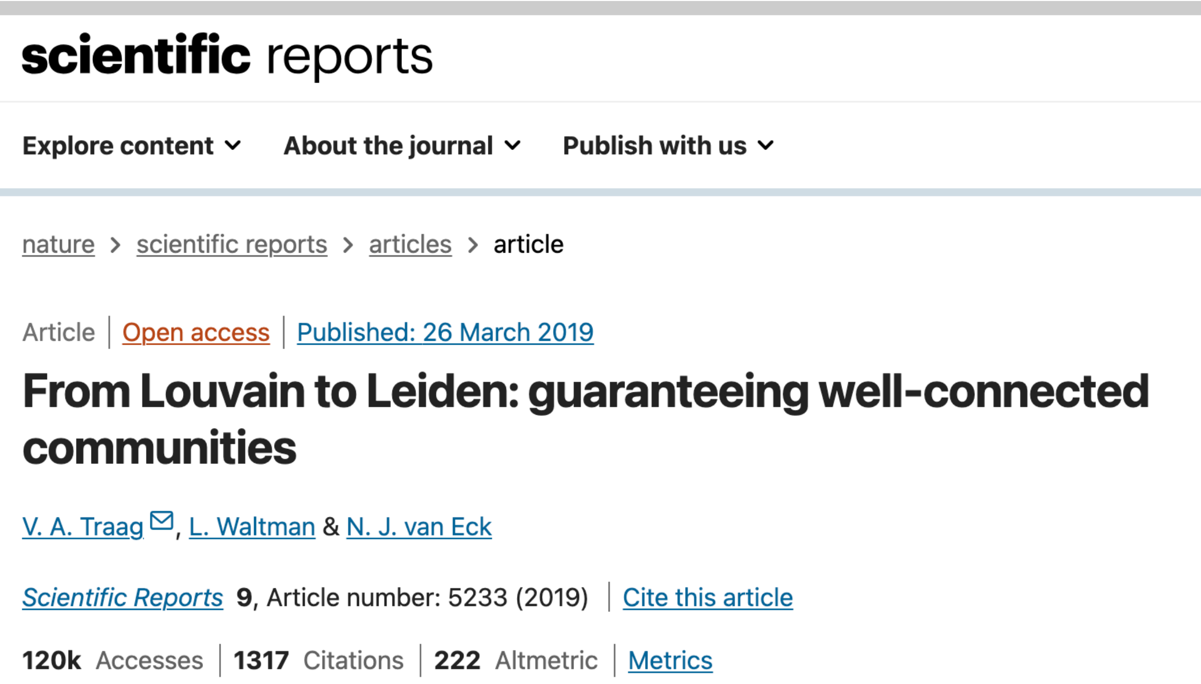
Community finding (aka “clustering”)

- Given a network (i.e., graph, with vertices and edges), partition the vertices into disjoint sets so that each set looks like a cohesive group.
- These groups are called “communities” or “clusters” or “modules” or “blocks”
- What features should communities have?
  - Dense (more edges inside than expected)
  - Separated from other communities
  - Connected, and even well-connected
  - Sometimes, a particular community size is sought

Well-connected = no small edge cut

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





Article | Open access | Published: 26 March 2019

From Louvain to Leiden: guaranteeing well-connected communities

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

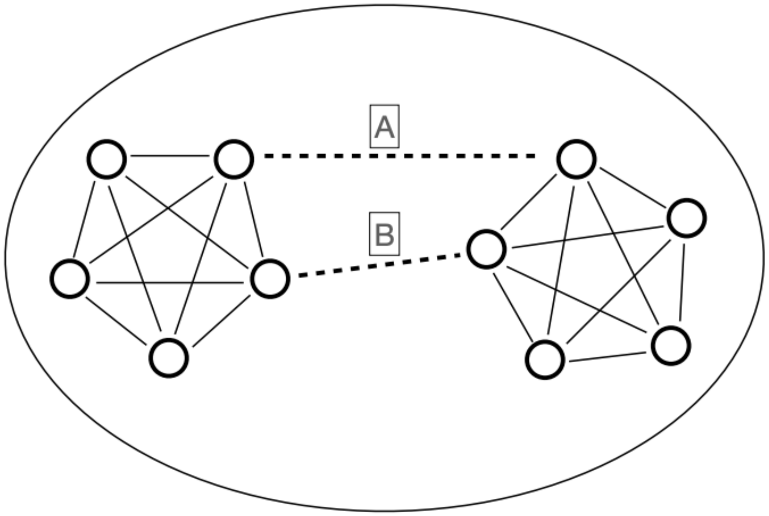
Scientific Reports **9**, Article number: 5233 (2019) | Cite this article

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

1. *Introduced Leiden algorithm*
2. *Demonstrates Louvain (for modularity) produces disconnected clusters*
3. *Proves that optimizing clustering under the Constant Potts Model is always "well-connected" (next slide)*
4. *Proves Leiden heuristic produces connected clusters*

Well-connected = no small edge cut

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



The CPM score and optimization problem

Given a network and a resolution parameter γ , find a partition of the nodes into disjoint clusters to maximize the CPM score

e_c is # edges in cluster c,

n_c is # nodes in cluster c

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

CPM-optimal clusterings are well-connected

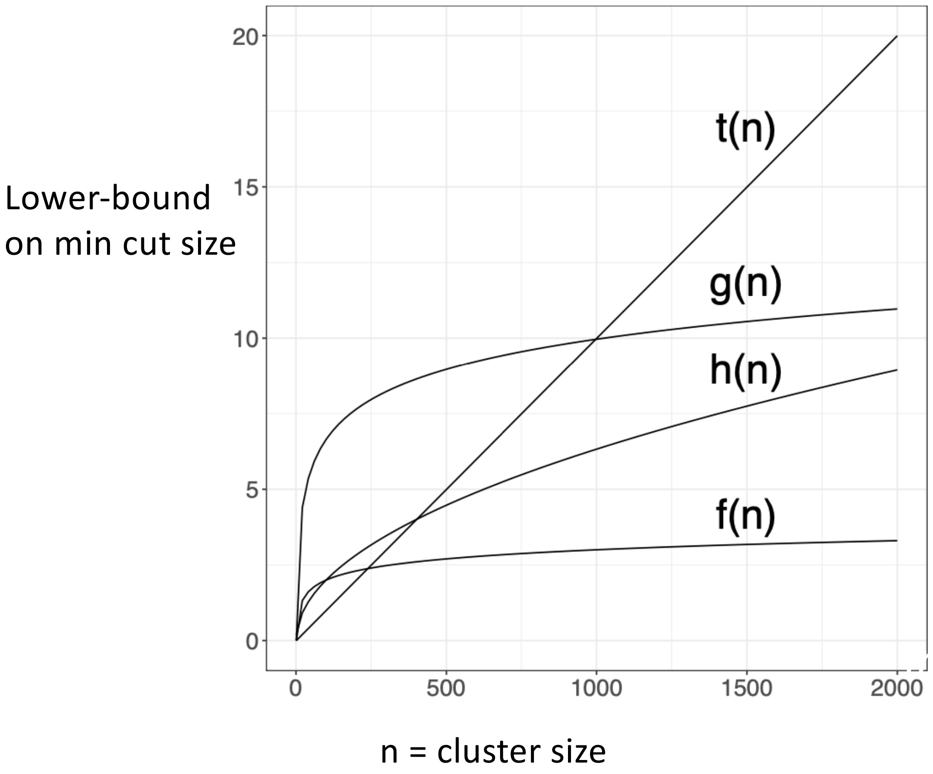
Recall: CPM optimization score depends on the resolution parameter γ

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

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

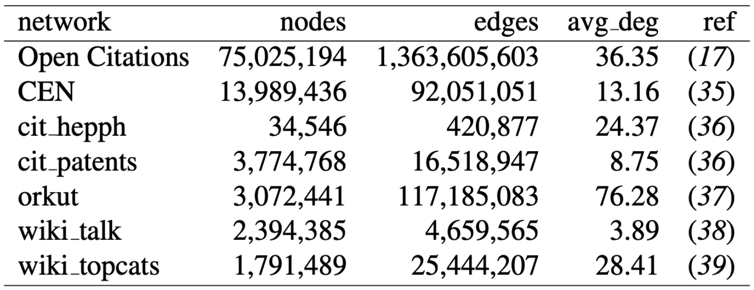
Lower bounds for “well-connected” clusters with n nodes



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

f(n) = $\log_{10} n$
g(n) = $\log_{2} n$
h(n) = $(n^{0.5})/5$

Park et al. (Complex Networks 2023): Well-Connected Communities in Real-World and Synthetic Networks



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.

Community Detection Methods:

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

M. Park*, Y. Tabatabaee*, V. Ramavarapu, B. Liu, V. Kamath Pailodi, R. Ramachandran, D. Korobskiy, F. Ayres, G. Chacko, and T. Warnow

Park et al. study results (preview)

- We demonstrate that all studied clustering methods produce clusters with small edge cuts on real world networks.
- Only Leiden and IKC completed on Open Citations.

- We present the Connectivity Modifier: flexible pipeline, modifies clustering to ensure well-connectivity, according to a user-provided rule.

Leiden clusters have small edge cuts

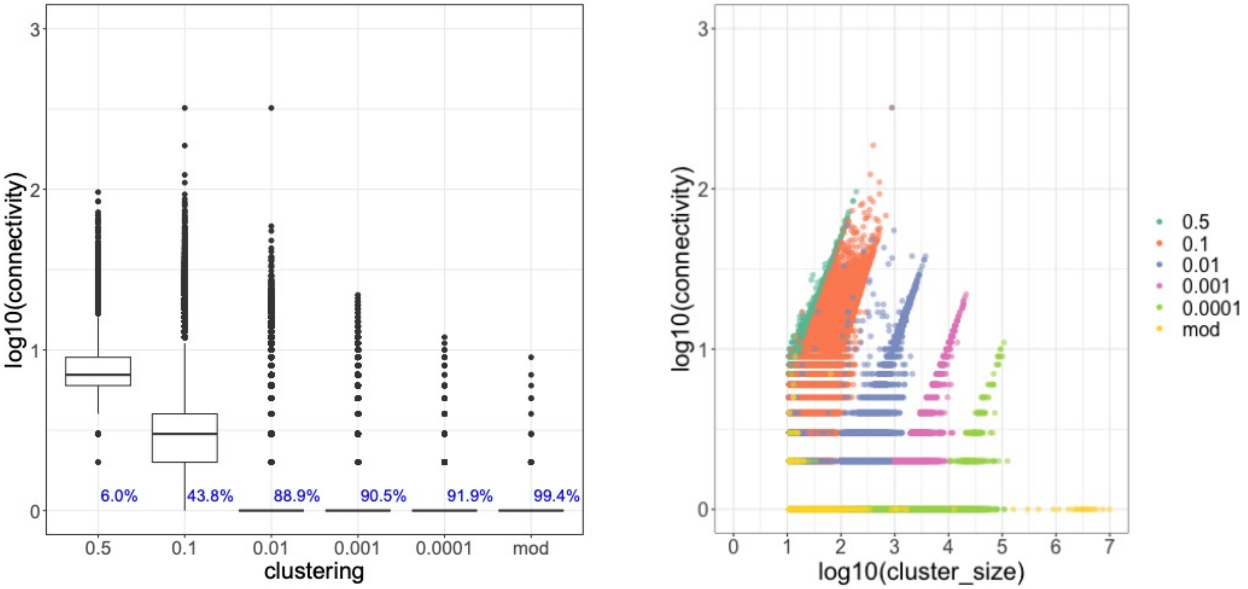


Figure 1: *Node coverage, connectivity, and size distribution of clusters generated by Leiden optimizing either CPM or modularity on the Open Citations network (75,025,194 nodes).*

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

- Blue text in left figure indicates node coverage

- Infomap, Markov Clustering, and Iterative k-core also produced clusters with small edge cuts.

Leiden clusters have small edge cuts

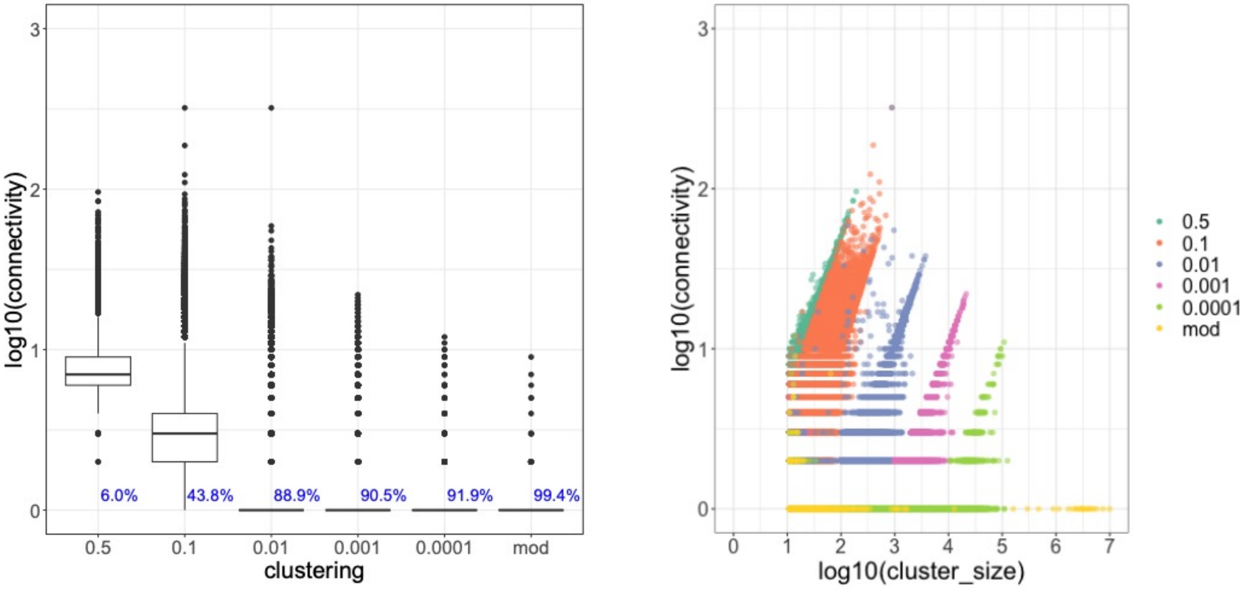


Figure 1: *Node coverage, connectivity, and size distribution of clusters generated by Leiden optimizing either CPM or modularity on the Open Citations network (75,025,194 nodes).*

- Only Leiden and IKC could complete on all networks we tested
- IKC had much lower node coverage than Leiden
- Conclusion: Trade-off between node coverage and edge-connectivity

CPM-optimal clusterings are well-connected

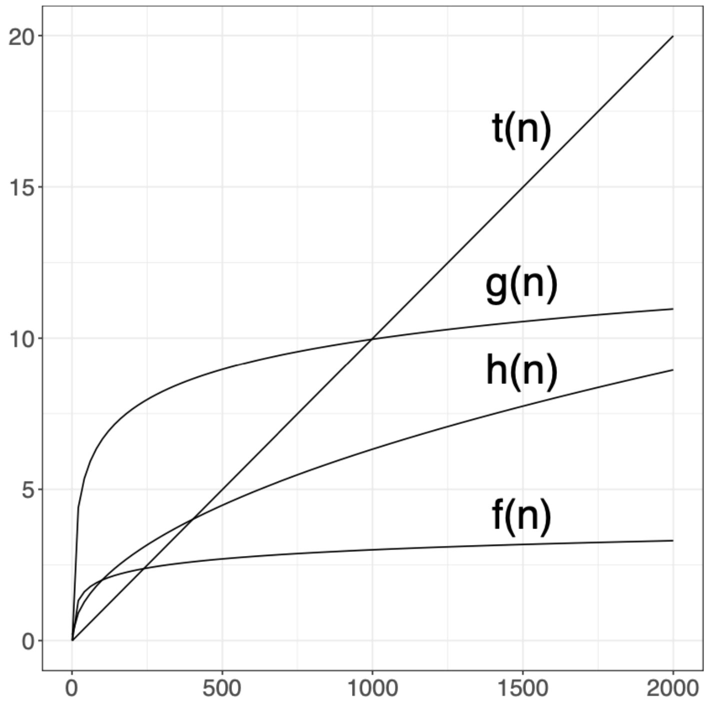
Recall: CPM optimization score depends on the resolution parameter γ

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

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

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

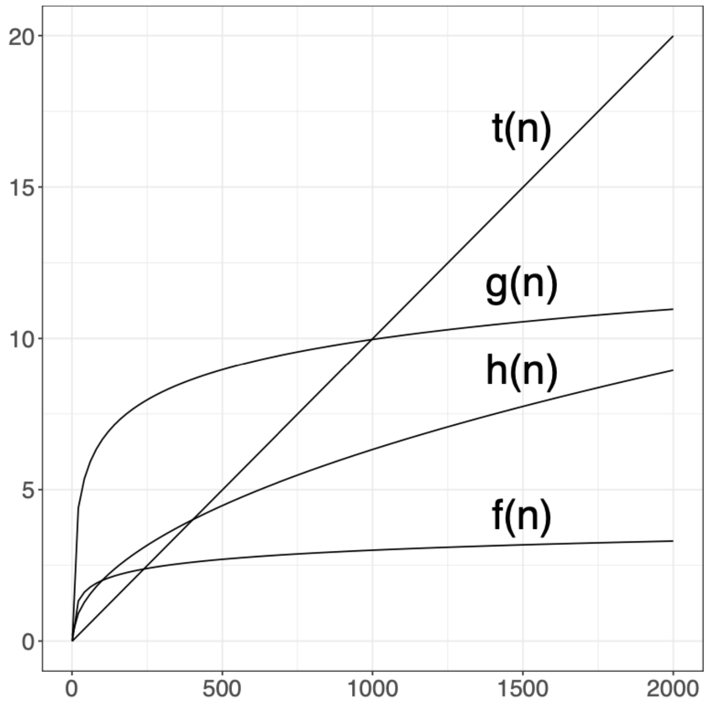


n = cluster size

$f(n) = \log_{10} n$
$g(n) = \log_2 n$
$h(n) = (n^{0.5})/5$

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

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



n = cluster size

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

We select f(n):

We consider a cluster with n nodes to be “well-connected” if the min-cut size exceeds f(n).

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

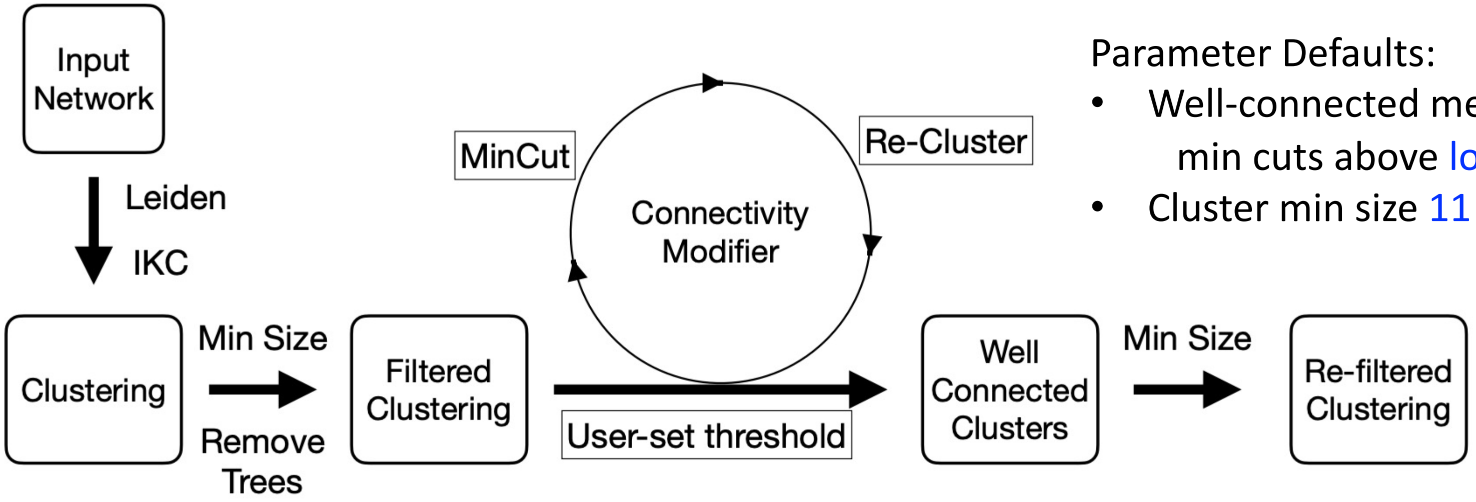
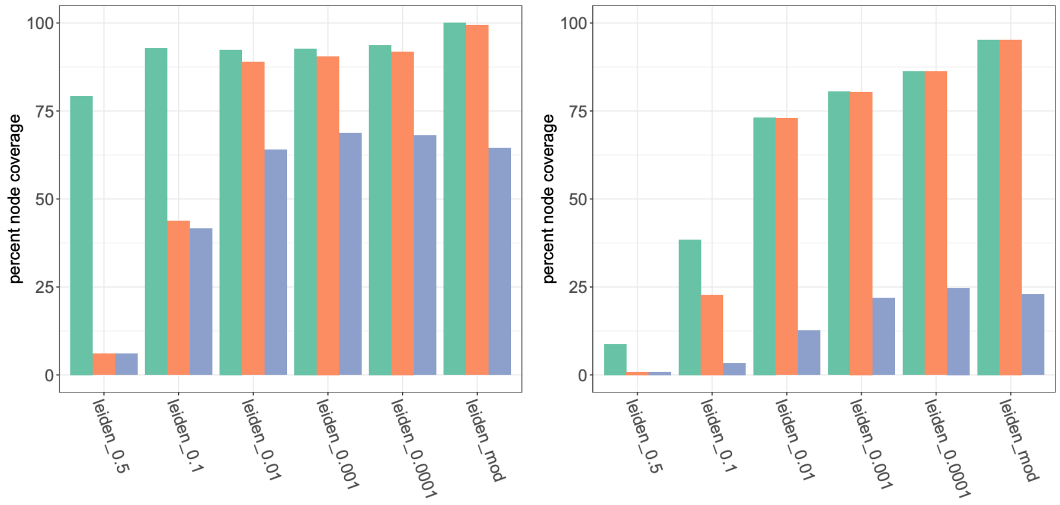


Figure 3: *Connectivity Modifier Pipeline Schematic.* The four-stage pipeline depends on user-

CM reduces node coverage



(a) Open Citations

(b) CEN

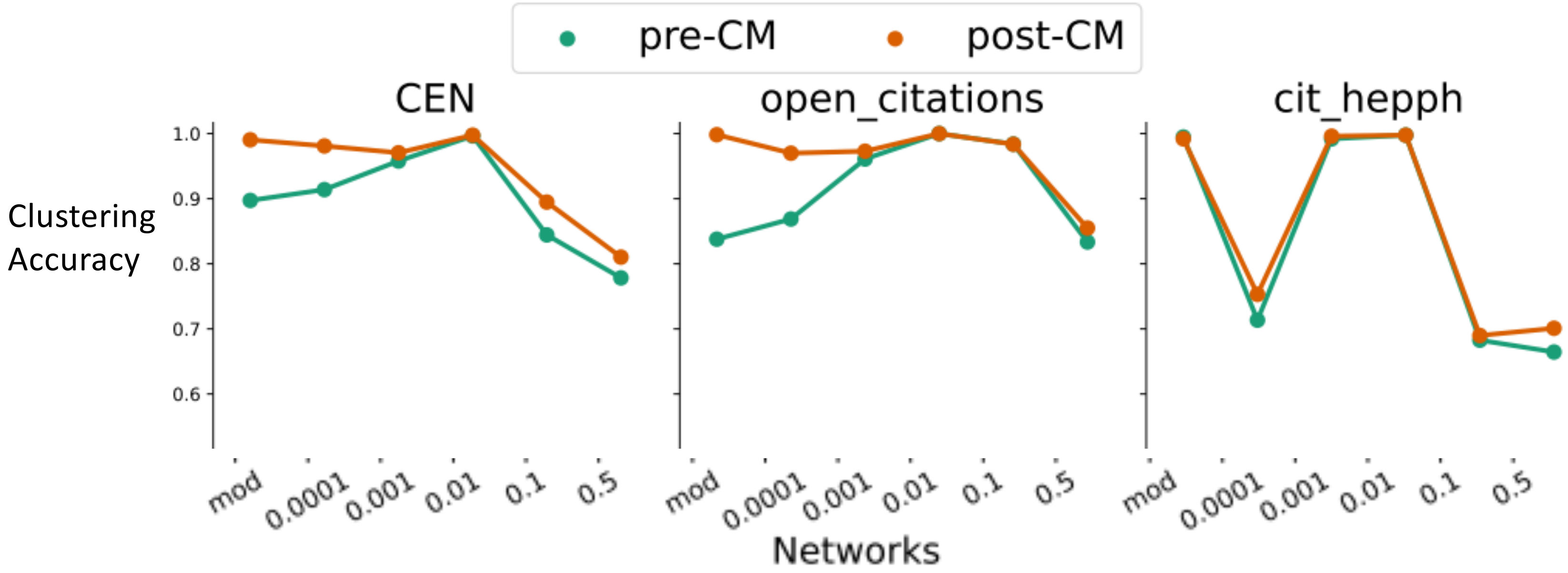
Figure 4: *Reduction in node coverage after CM treatment of Leiden clusters.* The Open Citations (left panel) and CEN (right panel) networks were clustered using the Leiden algorithm under CPM at five different resolution values or modularity. Node coverage (defined as the percentage of nodes in cluster 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

Tradeoff between node coverage and well-connectedness

Leiden-CPM and Leiden-Mod produce tree clusters

CM improves accuracy on synthetic networks



Results for NMI accuracy on LFR networks.
Results for other criteria and LFR networks are similar.

Observations, part 1

- For methods studied without CM post-processing, Leiden-CPM was the best of the tested methods (higher node coverage 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
- CM guarantees well-connectedness, but node coverage is substantially reduced by running CM

Additional Observations and Questions

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 Leiden-modularity or Leiden-CPM with small resolution parameter)

Additional Observations and Questions

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 Leiden-modularity or Leiden-CPM with small resolution parameter)

Perhaps these networks are not fully covered by communities?

Ongoing work (after the conference)

- We were asked at the conference if we had looked at Stochastic Block Models – we hadn’t at that time, so now we have!
- Participants:
  - PhD student Minhyuk Park
  - Undergraduates Daniel Feng and Siya Digra
- New (unpublished) results: SBMs also not great!

Stochastic Block Model: Clusters are often disconnected

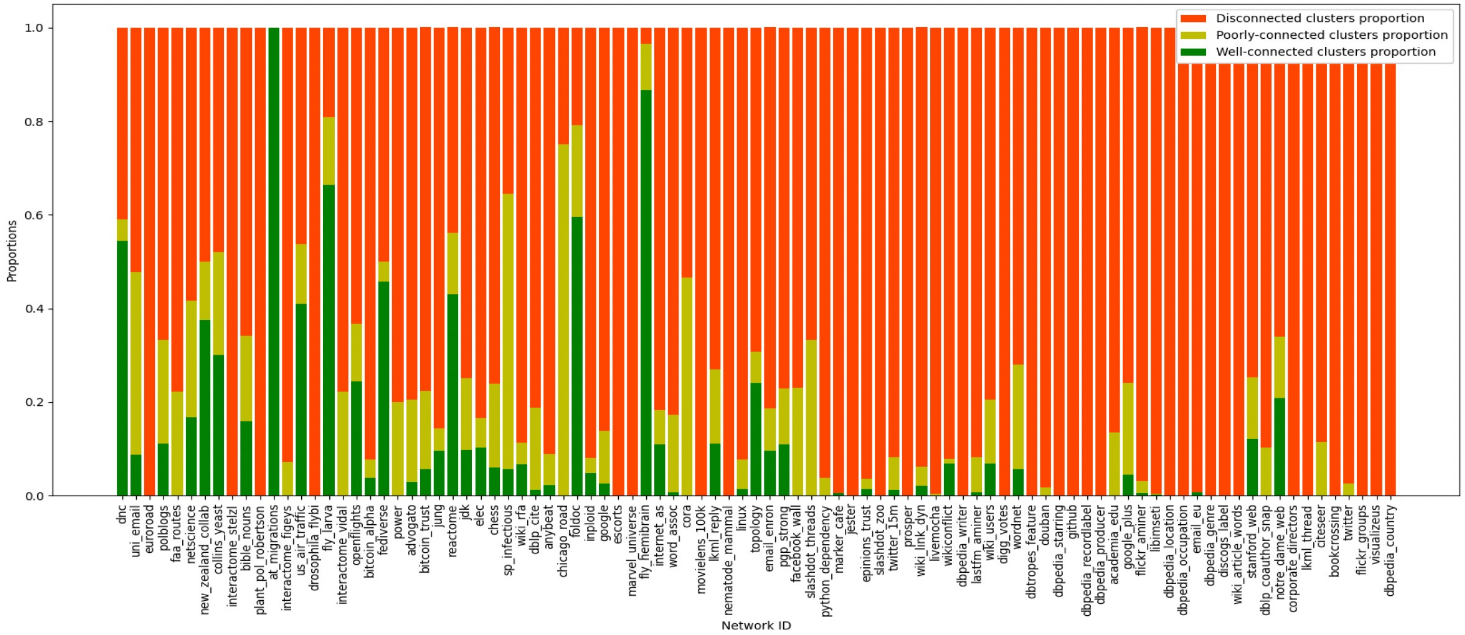


Figure 2: **Connectivity Proportions of DC SBM on Medium-Sized Peixoto Networks** The propor-

Red means disconnected
Light green: poorly connected
Dark green: well-connected

Networks have 1000-1,000,000 nodes, taken from Peixoto collection of networks

SBM run with degree-corrected model (similar results for other SBM models)

Stochastic Block Model: Clusters are often disconnected

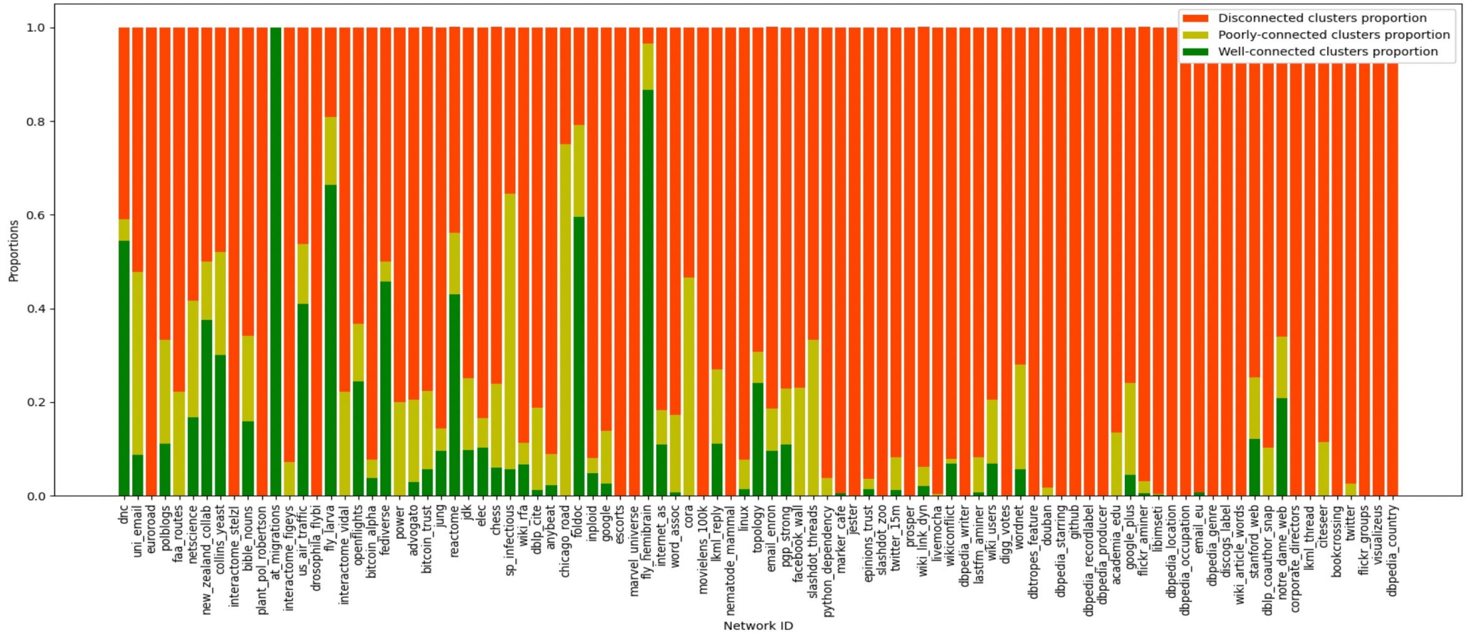


Figure 2: **Connectivity Proportions of DC SBM on Medium-Sized Peixoto Networks** The propor-

Red means disconnected
Light green: poorly connected
Dark green: well-connected

Networks have 1000-1,000,000 nodes, taken from Peixoto collection of networks

SBM run with degree-corrected model (similar results for other SBM models)

Following by CM ensures well-connectedness but reduces node coverage substantially (data not shown)

Take home points

- All tested clustering methods (Leiden-CPM, Leiden-modularity, Markov Clustering, Infomap, Stochastic Block Models, and Iterative k-core) produced clusters that had small edge cuts, and some produced disconnected clusters.
- The frequency and degree depends on the clustering method and network.
- The Connectivity Modifier (CM) provides a simple technique to ensure that all clusters are well-connected, but this reduces node coverage.

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:
  - Clusters should be checked for edge connectivity.
  - Ensuring edge-connectivity should be part of community detection methods.
  - The Connectivity Modifier can be used to improve clusterings.

Future work

- Developing other approaches for ensuring well-connectedness in communities
- Selecting threshold for well-connectedness based on network (instead of ad hoc, as done now)
- Evaluating other synthetic network simulators (e.g., SBM and ABCD)
- Developing improved simulators that come closer to real-world networks and clusterings

The CM code is open source

- CM is open source code (github) and under development, so that other clustering methods can be integrated.
- The algorithmic parameters (e.g., what "well-connected" means) can be modified.
- CM is fast enough to use on large networks.
- We welcome collaborations.
- See https://github.com/illinois-or-research-analytics/cm_pipeline
- See https://tandy.cs.illinois.edu/bibliometrics.html for full paper