Supplementary Materials for "Well-Connectedness and Community Detection"

Minhyuk Park*[1], Yasamin Tabatabaee*[1], Vikram Ramavarapu*[1], Baqiao Liu[1], Vidya Kamath Pailodi[1], Rajiv Ramachandran[1], Dmitriy Korobskiy[2], Fabio Ayres[3], George Chacko†[1,4,*], and Tandy Warnow‡[1,*]

[1]Siebel School of Computing and Data Science, University of Illinois Urbana-Champaign, Urbana, IL 61801, USA
[2]NTT DATA, McLean, VA 22102, USA
[3]Insper Institute, São Paulo, Brazil
[4]Office of Research, Grainger College of Engineering, University of Illinois Urbana-Champaign, Urbana, IL 61801, USA

Contents

*Minhyuk Park, Yasamin Tabatabaee, and Vikram Ramavarapu contributed equally
†chackoge@illinois.edu
‡warnow@illinois.edu

List of Tables

List of Figures

A Connectivity Modifier Pipeline

Preprocessing an input network

We remove duplicate/parallel edges and self-loops from the networks, using the `cleanup_el.R` script with the following command:

```
Rscript cleanup_el.R <original_network.tsv> <cleaned_network.tsv>
```

This step is now included as Stage 1 in the pipeline.

CM Pipeline

The cleaned input network $\mathcal{G}$ is then processed in a pipeline that has the following stages:

- **Clustering**: For each clustering method we used, we ran it in default mode. Specifically, for IKC (Iterative k-core) is run using $k = 10$ (which means that every output cluster has in-degree of at least k for every node) and for MCL we used inflation factor 2.0. For specific software, we used the Python version of the Leiden algorithm (leidenalg v0.8.2), Iterative K-core Clustering (IKC) method (v1.0.0), MCL (v22-282), and Infomap (v2.7.1). with the following commands respectively:

```
Leiden-CPM:
leidenalg.find_partition(net, leidenalg.CPMVertexPartition,
    resolution_parameter=r)
```

```
Leiden-Modularity:
leidenalg.find_partition(net, leidenalg.ModularityVertexPartition)
```

```
Iterative k-core (IKC):
IKC.py -e <cleaned_network.tsv> -k 10 -o <output file>
```

```
Markov Clustering (MCL):
mcl ${current_tsv_path} --abc -o ${output_file} -I 2.0
```

```
Infomap:
infomap ${current_tsv_path} ${output_path} --clu --ftree
```

  For the Open Citations network, we ran CM on the IKC clustering as follows. Each cluster of size greater than 100,000 nodes was run in a separate CM analysis until completion or failure. The remaining clusters were run in a single CM analysis together. The output clusterings from all of the individual CM runs were then combined to make the final output clustering. Two of the CM runs, originating from two large IKC clusters (#1637 and #2314), could not complete due to memory issues. Any nodes from these two clusters were returned as singleton clusters in the final output clustering. Thus, node coverage recorded was lower than if CM had completed on these two clusters.

- **Filtering during pre-processing**: Removing clusters of size at most 10 as well as trees (defined as acyclic connected clusters, or equivalently connected clusters where the number of edges is exactly one less than the number of nodes). We used the following command for this step:

  – `subset_graph_nonnetworkit.R` followed by `make_cm_ready.R`

- **CM**: Applying connectivity modifier with the following commands (assuming the clustering method used is Leiden), where the option `-g` specifies the resolution parameter. The version of CM used was v3.4.3 and for use with Infomap and MCL, was v4.0.1 (which provides support for these two methods):

```
$ cm -i <cleaned_network.tsv> -c leiden -e <filtered_leiden_clustering.tsv>
-g <r> -t 1log10 -o <cm_output.tsv>
$ cm2universal -g <cleaned_network.tsv> -i <cm_output.tsv> -o <cm_output.tsv>
$ python3 -m hm01.cm -i (network file) -e (clustering file) -t 1log10 -g 0.001
-c leiden -q -o out.tsv (versions 3.4.3 and 4.0.1)
```

- **Filtering during post-processing**: Removing clusters of size at most 10, using the same commands as the first filtering step. *Note:* A bug was discovered in April 2024 that affected a small number of clusters if the input file's encoding was not correctly detected as UTF-8. Based on our investigation, the bug only affected the memberships of the impacted clusters and did not change the final set of cluster IDs. The pre-corrected version of the script was used for the Leiden + CM analyses of the LFR networks and IKC+CM analyses of OC and CEN, and the corrected version of the script was used for the remainder of the study.
  - `post_cm_filter.R`: the default, a simple script that removes all clusters of size 10 or less.

The current version of the CM pipeline is available at:
https://github.com/illinois-or-research-analytics/cm_pipeline/

This version builds upon an earlier version, which is available at:
https://github.com/RuneBlaze/connectivity-modifier.

The current version of CM (v4.0.1) is more performant than the earlier version, improves upon how disconnected clusters are handled, whether given as input or created during the CM pipeline), and changes were made to ensure that 'extant clusters' were reported correctly.

B Additional Materials and Methods

B.1 Datasets

For all networks, we remove duplicate/parallel edges and self-loops from the networks, using the `cleanup_el.R` script with the following command:

```
Rscript cleanup_el.R <original_network.tsv> <cleaned_network.tsv>
```

B.2 Emulating real-world networks with LFR

In this experiment, we used Leiden optimizing modularity and CPM, with different resolution parameters, to produce clusterings of each of the seven real-world networks.

The code for estimating the parameters of a network/clustering and generating an LFR benchmark graph that represents their properties is available at https://github.com/ytabatabaee/emulate-real-nets and was used with the following commands.

```
python3 estimate_properties.py -n <network.tsv> -c <clustering_memberships.tsv>
```

that produces a json file containing all parameters of the network/clustering pair and then this json file is used to generate an LFR graph as follows

```
python3 gen_lfr.py -n <clustering_memberships.json> -lp <lfr-benchmark-software-
    path>
-cm <cmin_value>
```

in which the LFR software is used with the following command:

```
./binary_networks/benchmark -N <node-count> -k <avg-degree>
                            -maxk <max-degree> -mu <mixing-parameter>
                            -maxc <max-cluster-size> -minc <min-cluster-size>
                            -t1 <degree-exponent> -t2 <comm-size-exponent>
```

A somewhat limited range of parameters can be used to successfully generate an LFR graph. Due to restrictions of the LFR methodology, we made the following adjustments in our experiments for emulating our collection of real-world networks:

- LFR software is not scalable to large networks, and therefore we emulated the two largest networks (Open Citations and CEN) with LFR graphs with 3,000,000 nodes, and adjusted the maximum degree and maximum community size accordingly.
- The Wikipedia talk network (wiki_talk) has a very small average degree of 3.89, a maximum degree of 100029, and an estimated power-law degree exponent of 1.90. Based on the assumptions of the LFR methodology, we had to reduce the input maximum degree (`maxk`) to 31, as for the given average degree, any value above this for `maxk` resulted in an error from the LFR software.
- Singletons were always present in the clusterings of all the real-world networks; hence, the minimum community size was 1. However, the LFR methodology assumes that all nodes are in a valid community. In most cases, when setting `minc` $= 1$, the LFR software either took very long to run (more than our time limit of 4 hours per network), or could not generate the community size distribution with the given properties at all. Therefore, with the exception of the cit_hepph, a high energy physics citation network, for which `minc` $= 1$ successfully generated the LFR, for other networks, we found the *minimum* value of `minc` that generated the LFR graph in less than 4 hours with a brute-force search (exploring all values of `minc` starting from 1, to the point where the LFR was generated in the given time limit).

The full set of parameters we used for generating the LFR graphs based on the real-world networks are provided at `https://github.com/ytabatabaee/emulate-real-nets/tree/main/data_characteristics`, but see Tables K and L for some of these parameter values (e.g., mixing parameters).

The LFR software was unable to produce a network on wiki_talk with Leiden-CPM 0.5, on wiki_topcats with Leiden-CPM 0.5, or on orkut with any of the Leiden clusterings.

B.3 Parameter Exploration for MCL

The default setting for the inflation parameter for MCL, according to the guidelines published by the developer, is 2.0. Here we explore MCL under various inflation factors in the range of 1.0 to 5.0 on the cit_hepph network as shown in Table A. Using inflation factor 1.0 produces only 61 clusters, compared to the other clusterings that ranged from 3036 clusters to 6060 clusters. Furthermore, using inflation factor 1.0 produced 0% well-connected clusters of size at least 11, and also produces a very large cluster with 34,401 nodes, making this choice of the inflation factor a poor one.

After removing 1.0 from consideration, we now consider the remaining choices for the inflation factor. MCL using the other inflation parameter values produced relatively similar percentages of disconnected clusters, in the range of 6-15% for the clusters of size at least 11, and had nearly identical minimum and median cluster sizes. They also had fairly similar percentages of clusters of size at least 11 that were well-connected. However, when restricting to clusters of size at least 11, they differed substantially in node coverage, with the choice of 2.0 producing having 70% of the nodes in clusters and the other choices having 8% to 28%. Thus, the default inflation parameter of 2.0 yields better results in terms of node coverage than the other choices, and similar results with respect to the other criteria.

Table A: Impact of varying inflation parameter of MCL on cit_hepph real-world network We show the inflation factor for MCL in the first column along with the number of clusters, the cluster size and the distributions of the normalized mincut size (i.e., mincut size divided by $\log_{10}(n)$ where n is the cluster size) in min/median/max format, node coverage (expressed as a proportion) of the nodes in clusters of size at least 2 or of size at least 11 (last column). The network cit_hepph has 34,546 nodes. Note that when the normalized mincut size is 0, this indicates that the cluster is not connected. When the normalized mincut size is strictly greater than 1.0, then the cluster is considered well-connected, and otherwise it is considered poorly connected.

inflation factor	# clusters	Cluster size dist. min/median/max	% disconn. $n \geq 2/n \geq 11$	% well.-conn. $n \geq 2/n \geq 11$	node coverage $n \geq 2/n \geq 11$
1.0	61	2/2/34401	0/0	98/0	1.00/1.00
2.0	3036	2/4/887	7/13	81/29	0.97/0.70
3.0	5728	2/3/518	7/15	90/34	0.79/0.28
4.0	6060	2/2/341	5/11	93/31	0.65/0.14
5.0	5884	2/2/264	4/6	95/32	0.56/0.08

B.4 Parameter exploration for Infomap

We explored Infomap under different numbers of trials on the six real-world networks on which it completed, to determine if five trials was better than on trial in terms of percentage of clusters of size at least 11 that were well-connected and percentage disconnected, and node coverage from these clusters. Results, shown in Table B, show that changing the number of trials had overall no benefit to Infomap with respect to these statistics. Specifically, the number of trials did not change the node coverage at all (100% for both one and five trials on all networks); it also did not change the percent of clusters of size at least 11 that were well-connected, and these were uniformly very low (0% for five of the network and 3% for the remaining one). The percent disconnected decreased for one network (the smallest, cit_hepph), was unchanged for two, and increased for three.

Table B: Impact of number of trials on Infomap Results of running Infomap for one and five trials are shown for six real-world networks (rows). For each number of trials, the number of clusters of size at least two and cluster size distribution are shown. For clusters of size at least 2 or of size at least 11, the node coverage, percent disconnected, and percent well-connected are shown, each given as x/y, where x is for clusters of size at least 2 and y is for clusters of size at least 11

	Num Clusters		Cluster Size Dist min/median/max		Node Coverage		% disconn. $n \geq 2/n \geq 11$		% well.-conn.	
	1	5	1	5	1	5	1	5	1	5
cit_hepph	66	66	2/2/13862	2/2/14607	1.0/1.0	1.0/1.0	3/29	2/17	89/0	91/0
cit_patents	3902	3919	2/2/245582	2/2/243548	1.0/1.0	1.0/1.0	3/44	4/46	92/0	92/0
orkut	20	15	8/26660/1417010	41/64835/1417071	1.0/1.0	1.0/1.0	75/79	80/80	5/0	0/0
wiki_talk	22191	22227	2/17/53773	2/17/53773	1.0/1.0	1.0/1.0	0/0	0/0	35/0	35/0
wiki_topcats	8837	8332	2/14/79644	2/14/97346	1.0/1.0	1.0/1.0	5/7	5/7	37/3	38/3
cen	187	197	2/103/2314354	2/154/2363727	1.0/1.0	1.0/1.0	4/6	6/8	30/0	29/0

B.5 Exploring LGI-MCL

We also explored LGI-MCL, using four different kernels, on the cit_hepph graph. The kernels were "EBC", "RA', "EBC-RA", and "original". LGI-MCL requires a target number of clusters; we therefore ran MCL in default mode and obtained a clustering, and gave LGI-MCL the number of clusters in this clustering as the target. We evaluated each output clustering in terms of its node coverage, cluster size distribution, percent of clusters disconnected, and percent of clusters well-connected; all statistics were reported for clusters of size at least two or of size at least 11.

Results on this experiment are shown in Table C. The four LGI-MCL variants have very similar results with respect to the criteria we considered. After restricting to clusters of size at least 11, MCL has higher node coverage (69% compared to 55%-57%) and percentage of disconnected clusters (13% compared to 10% or 11%). Furthermore, MCL has a much larger percentage of well-connected clusters of size at least 11 than the LGI-MCL variants (29% compared to 8–10%).

The original LGI-MCL implementation can be found at https://github.com/biomedical-cybernetics/LGI-MCL/tree/main while our minimal modifications to make the code compatible with linux cluster environments used to test LGI-MCL can be found at https://github.com/MinhyukPark/LGI-MCL/tree/main/linux_code. This forked repository contains compilation instructions and commands used for running these experiments.

```
./modified_mcl <input edgelist> <kernel> <output prefix>
```

Table C: Comparison of LGI-MCL and MCL on cit_hepph real-world network MCL and various kernels of LGI-MCL were used to cluster the cit_hepph real-world network. All runs of LGI-MCL were given 3036, the number of clusters returned from MCL with inflation factor 2.0, as its target number of clusters. Node coverage columns show proportion of nodes in clusters of size at least two, or of size at least 11. The distributions shown for cluster size are restricted to clusters of size at least two.

	# clusters	node cov. $n \geq 2$ / $n \geq 11$	Cluster Size dist. min/median/max	% disconn. $n \geq 2$ / $n \geq 11$	% well-conn. $n \geq 2$ / $n \geq 11$
MCL	3036	0.98/0.69	2/4/887	7/13	81/29
LGI-MCL(original)	3817	0.93/0.57	2/4/909	3/11	83/8
LGI-MCL(RA)	4117	0.94/0.55	2/4/885	3/11	85/9
LGI-MCL(EBC)	3742	0.93/0.57	2/4/879	2/10	83/9
LGI-MCL(EBC_RA)	4076	0.93/0.55	2/4/861	3/11	85/10

C Additional Results on Real-World Networks

C.1 Properties of clusterings on real-world networks

Table D: Cluster count and size information for clusterings of real-world networks. For each clustering of each of the real-world networks, the number of clusters of size at least 2 and 11, and the distribution of cluster sizes are shown.

network	method	parameter	clus_count $n \geq 2$	% nodes in clusters $n \geq 11$.	clust. size min	clust size median	clust size max
oc	leiden	0.5	20966119	6.0	2	2	192
oc	leiden	0.1	8642175	43.8	2	5	882
oc	leiden	0.01	2134603	88.9	2	15	3729
oc	leiden	0.001	839902	90.5	2	3	21385
oc	leiden	0.0001	561116	91.9	2	2	125017
oc	leiden	modularity	184257	99.4	2	2	9922297
oc	ikc	10	2569	23.6	11	40	6650349
cen	leiden	0.5	433761	1.0	2	2	70
cen	leiden	0.1	517669	24.0	2	11	320
cen	leiden	0.01	280544	76.6	2	28	3236
cen	leiden	0.001	66470	84.5	2	97	13025
cen	leiden	0.0001	11779	90.5	2	377	70138
cen	leiden	modularity	253	100.0	2	28	1312837
cen	ikc	10	128	3.8	14	79	214877
cen	infomap	default	187	100.0	2	103	2314354
cit_hepph	leiden	0.5	7569	16.3	2	3	55
cit_hepph	leiden	0.1	3050	63.2	2	6	169
cit_hepph	leiden	0.01	810	91.5	2	9	744
cit_hepph	leiden	0.001	366	94.3	2	3	3436
cit_hepph	leiden	0.0001	284	95.8	2	2	20000
cit_hepph	leiden	modularity	82	99.5	2	2	4031
cit_hepph	ikc	10	28	20.1	14	59.5	1530
cit_hepph	infomap	default	68	99.6	2	2	13027
cit_hepph	mcl	2.0	3036	69.0	2	4	887
cit_patents	leiden	0.5	1143221	1.1	2	2	116
cit_patents	leiden	0.1	556688	26.6	2	5	330
cit_patents	leiden	0.01	134380	95.5	2	19	804
cit_patents	leiden	0.001	29463	98.7	2	63	3268
cit_patents	leiden	0.0001	9125	99.3	2	4	20000
cit_patents	leiden	modularity	3708	99.7	2	2	198537
cit_patents	ikc	10	582	2.0	15	35	23468
cit_patents	infomap	default	3902	99.7	2	2	245582
orkut	leiden	0.5	617430	30.2	2	2	217
orkut	leiden	0.1	232145	64.8	2	6	1444
orkut	leiden	0.01	47094	95.5	2	21	12016
orkut	leiden	0.001	11269	96.1	2	44	46062
orkut	leiden	0.0001	11348	94.1	2	2	190195
orkut	leiden	modularity	36	100.0	3	22572.5	800146
orkut	ikc	10	758	43.7	11	29	386103
orkut	infomap	default	20	100.0	8	26660	1417010
wiki_talk	leiden	0.5	47808	0.1	2	2	133
wiki_talk	leiden	0.1	48481	0.7	2	4	1141
wiki_talk	leiden	0.01	31983	31.9	2	11	200
wiki_talk	leiden	0.001	16035	79.3	2	48	1205
wiki_talk	leiden	0.0001	7315	97.5	2	167	13635
wiki_talk	leiden	modularity	2739	99.8	2	2	225817
wiki_talk	ikc	NA	NA	NA	NA	NA	NA
wiki_topcats	leiden	0.5	455139	2.2	2	2	89
wiki_topcats	leiden	0.1	240673	29.3	2	5	523
wiki_topcats	leiden	0.01	58343	91.4	2	17	2547
wiki_topcats	leiden	0.001	20257	93.3	2	3	18024
wiki_topcats	leiden	0.0001	13269	94.5	2	2	125725
wiki_topcats	leiden	modularity	27	100.0	155	45467	273395
wiki_topcats	ikc	10	170	6.8	11	28	45613
wiki_talk	infomap	default	22191	98.6	2	17	53773

Table E: Node coverage and cluster counts of clusterings of the 7 real-world networks .

network	method .	parameter	# clust. $11 \leq n$	# clust. $2 \leq n \leq 10$	# singletons	node cov. [%] $n \geq 11$	node cov. [%] $2 \leq n \leq 10$	node cov. [%] singletons	# nodes in clusters non-singletons	node cov. [%] $n \geq 2$
cen	ikc	10	128	0	13454271	3.8	0	96.2	535165	3.8
wiki_topcats	ikc	10	170	0	1669262	6.8	0	93.2	122227	6.8
cit_hepph	ikc	10	28	0	27611	20.1	0	79.9	6935	20.1
cit_patents	ikc	10	582	0	3699025	2	0	98	75743	2
orkut	ikc	10	758	0	1729355	43.7	0	56.3	1343086	43.7
oc	ikc	10	2569	0	57301796	23.6	0	76.4	17723398	23.6
cit_hepph	mcl	2.0	612	2424	862	69	28.5	2.5	33684	97.5
cen	infomap	default	130	57	0	100	0	0	13989436	100
wiki_talk	infomap	default	14491	7700	35	98.6	1.4	0	2394350	100
wiki_topcats	infomap	default	5676	3161	1	98.9	1.1	0	1791488	100
cit_hepph	infomap	default	8	60	0	99.6	0.4	0	34546	100
cit_patents	infomap	default	297	3605	0	99.7	0.3	0	3774768	100
orkut	infomap	default	19	1	0	100	0	0	3072441	100
cen	leiden	0.0001	11327	452	1323426	90.5	0	9.5	12666010	90.5
cen	leiden	0.001	65915	555	2165097	84.5	0	15.5	11824339	84.5
cen	leiden	0.01	275807	4737	3238339	76.6	0.3	23.1	10751097	76.9
cen	leiden	0.1	273971	243698	8332294	24	16.5	59.6	5657142	40.4
cen	leiden	0.5	8448	425313	12701688	1	8.2	90.8	1287748	9.2
cen	leiden	modularity	187	66	0	100	0	0	13989436	100
wiki_talk	leiden	0.0001	4488	2827	52664	97.5	0.3	2.2	2341721	97.8
wiki_talk	leiden	0.001	9622	6413	479532	79.3	0.6	20	1914853	80
wiki_talk	leiden	0.01	16278	15705	1585195	31.9	1.9	66.2	809190	33.8
wiki_talk	leiden	0.1	1403	47078	2145665	0.7	9.7	89.6	248720	10.4
wiki_talk	leiden	0.5	60	47748	2292833	0.1	4.2	95.8	101552	4.2
wiki_talk	leiden	modularity	166	2573	0	99.8	0.2	0	2394385	100
wiki_topcats	leiden	0.0001	1148	12121	64237	94.5	1.9	3.6	1727252	96.4
wiki_topcats	leiden	0.001	8014	12243	87381	93.3	1.9	4.9	1704108	95.1
wiki_topcats	leiden	0.01	44040	14303	96200	91.4	3.2	5.4	1695289	94.6
wiki_topcats	leiden	0.1	27734	212939	172523	29.3	61.1	9.6	1618966	90.4
wiki_topcats	leiden	0.5	2666	452473	575060	2.2	65.7	32.1	1216429	67.9
wiki_topcats	leiden	modularity	27	0	0	100	0	0	1791489	100
cit_hepph	leiden	0.0001	14	270	630	95.8	2.3	1.8	33916	98.2
cit_hepph	leiden	0.001	63	303	1029	94.3	2.7	3	33517	97
cit_hepph	leiden	0.01	380	430	1370	91.5	4.6	4	33176	96
cit_hepph	leiden	0.1	821	2229	2094	63.2	30.7	6.1	32452	93.9
cit_hepph	leiden	0.5	346	7223	5187	16.3	68.7	15	29359	85
cit_hepph	leiden	modularity	18	64	0	99.5	0.5	0	34546	100
cit_patents	leiden	0.0001	3238	5887	7621	99.3	0.5	0.2	3767147	99.8
cit_patents	leiden	0.001	21596	7867	21815	98.7	0.7	0.6	3752953	99.4
cit_patents	leiden	0.01	114308	20072	64155	95.5	2.8	1.7	3710613	98.3
cit_patents	leiden	0.1	58858	497830	123020	26.6	70.2	3.3	3651748	96.7
cit_patents	leiden	0.5	2653	1140568	1049310	1.1	71.1	27.8	2725458	72.2
cit_patents	leiden	modularity	94	3614	0	99.7	0.3	0	3774768	100
orkut	leiden	0.0001	691	10657	153956	94.1	0.9	5	2918485	95
orkut	leiden	0.001	6279	4990	106634	96.1	0.4	3.5	2965807	96.5
orkut	leiden	0.01	41535	5559	115873	95.5	0.7	3.8	2956568	96.2
orkut	leiden	0.1	59555	172590	147860	64.8	30.4	4.8	2924581	95.2
orkut	leiden	0.5	45170	572260	347743	30.2	58.5	11.3	2724698	88.7
orkut	leiden	modularity	34	2	0	100	0	0	3072441	100
oc	leiden	0.0001	38701	522415	4697754	91.9	1.8	6.3	70327440	93.7
oc	leiden	0.001	231899	608003	5487615	90.5	2.1	7.3	69537579	92.7
oc	leiden	0.01	1361462	773141	5721905	88.9	3.4	7.6	69303289	92.4
oc	leiden	0.1	1311357	7330818	5413061	43.8	49	7.2	69612133	92.8
oc	leiden	0.5	297230	20668889	15583737	6	73.2	20.8	59441457	79.2
oc	leiden	modularity	2184	182073	0	99.4	0.6	0	75025194	100

Table F: Tree Cluster counts using Leiden clustering on real-world networks: cen, cit_hepph, cit_patents, and open citations. The count of clusters of size at least 11 is provided for trees, stars (a type of tree), and non-tree clusters, for Leiden clustering optimizing either modularity or CPM. The total number of clusters of size at least 11 is the sum of these three types.

Network	Opt. criterion "Mod' or res. param for CPM	Cluster Type	Number clusters
cen	0.0001	non_tree	9520
cen	0.0001	star	1186
cen	0.0001	tree	621
cen	0.001	non_tree	43083
cen	0.001	star	18064
cen	0.001	tree	4768
cen	0.01	non_tree	64566
cen	0.01	star	205229
cen	0.01	tree	6012
cen	0.1	non_tree	18576
cen	0.1	star	255395
cen	0.5	non_tree	8448
cen	mod	non_tree	48
cen	mod	tree	8
cen	mod	star	131
cit_hepph	0.0001	non_tree	14
cit_hepph	0.001	non_tree	63
cit_hepph	0.01	non_tree	379
cit_hepph	0.01	tree	1
cit_hepph	0.1	non_tree	820
cit_hepph	0.1	star	1
cit_hepph	0.5	non_tree	346
cit_hepph	mod	non_tree	18
cit_patents	0.0001	non_tree	3024
cit_patents	0.0001	tree	178
cit_patents	0.0001	star	36
cit_patents	0.001	non_tree	20608
cit_patents	0.001	tree	939
cit_patents	0.001	star	49
cit_patents	0.01	non_tree	91410
cit_patents	0.01	star	415
cit_patents	0.01	tree	22483
cit_patents	0.1	non_tree	58543
cit_patents	0.1	star	315
cit_patents	0.5	non_tree	2653
cit_patents	mod	non_tree	75
cit_patents	mod	tree	9
cit_patents	mod	star	10
oc	0.0001	non_tree	33705
oc	0.0001	tree	3519
oc	0.0001	star	1477
oc	0.001	non_tree	207713
oc	0.001	star	1499
oc	0.001	tree	22687
oc	0.01	non_tree	1033761
oc	0.01	star	8591
oc	0.01	tree	319110
oc	0.1	non_tree	1297088
oc	0.1	star	14268
oc	0.1	tree	1
oc	0.5	non_tree	297230
oc	mod	non_tree	912
oc	mod	star	508
oc	mod	tree	764

Table G: Tree Cluster counts using Leiden clustering on orkut, wiki_talk, and wiki_topcats. The count of clusters of size at least 11 is provided for trees, stars (a type of tree), and non-tree clusters. The total number of clusters of size at least 11 is the sum of these three types.

Network	Opt. criterion "Mod' or res. param for CPM	Cluster Type	Number clusters
orkut	0.0001	non_tree	677
orkut	0.0001	tree	13
orkut	0.0001	star	1
orkut	0.001	non_tree	6199
orkut	0.001	tree	79
orkut	0.001	star	1
orkut	0.01	non_tree	37470
orkut	0.01	star	13
orkut	0.01	tree	4052
orkut	0.1	non_tree	59372
orkut	0.1	star	183
orkut	0.5	non_tree	45170
orkut	mod	non_tree	34
wiki_talk	0.0001	non_tree	3104
wiki_talk	0.0001	star	128
wiki_talk	0.0001	tree	1256
wiki_talk	0.001	non_tree	3760
wiki_talk	0.001	star	1797
wiki_talk	0.001	tree	4065
wiki_talk	0.01	non_tree	2579
wiki_talk	0.01	star	7850
wiki_talk	0.01	tree	5849
wiki_talk	0.1	non_tree	133
wiki_talk	0.1	star	1270
wiki_talk	0.5	non_tree	60
wiki_talk	mod	non_tree	149
wiki_talk	mod	tree	7
wiki_talk	mod	star	10
wiki_topcats	0.0001	non_tree	1088
wiki_topcats	0.0001	tree	47
wiki_topcats	0.0001	star	13
wiki_topcats	0.001	non_tree	7611
wiki_topcats	0.001	tree	311
wiki_topcats	0.001	star	92
wiki_topcats	0.01	non_tree	36854
wiki_topcats	0.01	star	891
wiki_topcats	0.01	tree	6295
wiki_topcats	0.1	non_tree	25364
wiki_topcats	0.1	star	2370
wiki_topcats	0.5	non_tree	2666
wiki_topcats	mod	non_tree	27

C.2 Connectivity of clusterings on real-world networks

Table H: Statistics of IKC, MCL, and Infomap clusterings of real-world networks The number of clusters and the cluster size distributions (minimum/median/maximum) are for all non-singleton clusters. Also shown are percent disconnected, percent well-connected clusters of size at least 2 or those at least 11.

Method	Network	# clus.	clus. size dist. min/median/max	% disconn. $n \geq 2$ / $n \geq 11$	% well-conn. $n \geq 2$ / $n \geq 11$
Infomap	CEN	187	2/103/2314354	4/6	30/0
Infomap	cit_hepph	68	2/2/13027	0/0	88/0
Infomap	cit_patents	3902	2/2/245582	3/44	92/0
Infomap	wiki_talk	22191	2/17/53773	0/0	35/0
Infomap	wiki_topcats	8837	2/14/79644	5/7	37/3
Infomap	orkut	20	8/26660/1417010	75/79	5/0
MCL	cit_hepph	3306	2/4/887	7/13	81/29
IKC	CEN	128	14/79/214877	0/0	86/86
IKC	OC	2569	11/40/6650349	0/0	94/94
IKC	cit_hepph	28	14/60/1530	0/0	93/93
IKC	cit_patents	582	15/35/23468	0/0	93/93
IKC	wiki_topcats	170	11/28/45613	0/0	92/92
IKC	orkut	758	11/29/386103	0/0	88/88

Table I: Node coverage (NC) before and after CM-processing for Leiden-CPM clusterings on real-world networks

method	**network**	**pre-CM NC** $n \geq 2$	**pre-CM NC** $n \geq 11$	**post-CM NC** $n \geq 11$
Leiden-CPM(0.0001)	CEN	0.905	0.905	0.259
Leiden-CPM(0.0001)	OC	0.937	0.919	0.680
Leiden-CPM(0.0001)	cit_hepph	0.982	0.958	0.835
Leiden-CPM(0.0001)	cit_patents	0.998	0.993	0.503
Leiden-CPM(0.0001)	wiki_talk	0.978	0.975	0.017
Leiden-CPM(0.0001)	wiki_topcats	0.964	0.945	0.695
Leiden-CPM(0.0001)	orkut	0.950	0.941	0.849
Leiden-CPM(0.001)	CEN	0.845	0.845	0.230
Leiden-CPM(0.001)	OC	0.927	0.905	0.687
Leiden-CPM(0.001)	cit_hepph	0.970	0.943	0.844
Leiden-CPM(0.001)	cit_patents	0.994	0.987	0.575
Leiden-CPM(0.001)	wiki_talk	0.800	0.793	0.008
Leiden-CPM(0.001)	wiki_topcats	0.951	0.933	0.638
Leiden-CPM(0.001)	orkut	0.965	0.961	0.833
Leiden-CPM(0.1)	CEN	0.404	0.240	0.036
Leiden-CPM(0.1)	OC	0.928	0.438	0.416
Leiden-CPM(0.1)	cit_hepph	0.939	0.632	0.612
Leiden-CPM(0.1)	cit_patents	0.967	0.266	0.238
Leiden-CPM(0.1)	wiki_talk	0.104	0.007	0.001
Leiden-CPM(0.1)	wiki_topcats	0.904	0.293	0.252
Leiden-CPM(0.1)	orkut	0.952	0.648	0.635
Leiden-CPM(0.5)	CEN	0.092	0.010	0.010
Leiden-CPM(0.5)	OC	0.792	0.060	0.060
Leiden-CPM(0.5)	cit_hepph	0.850	0.163	0.163
Leiden-CPM(0.5)	cit_patents	0.722	0.011	0.011
Leiden-CPM(0.5)	wiki_talk	0.042	0.001	0.001
Leiden-CPM(0.5)	wiki_topcats	0.679	0.022	0.022
Leiden-CPM(0.5)	orkut	0.887	0.302	0.302

Table J: Node coverage by stage for Leiden clusterings on real-world networks "% disc." and "% poor" refer to the percent of the clusters of size at least 2 that are disconnected or poorly connected, respectively. Leiden nc2, filter nc2, and cm nc2 refer to the node coverage (i.e., proportion of the node set that are in clusters) of size at least 2 in the input Leiden clustering, after Stage 2 (filtering), and after Stage 3 of the CM-pipeline.

	% disc.	% poor	Leiden nc2	Stage 2 nc2	Stage 3 nc2
cit hepph Leiden-modularity	0	23	1.00	0.99	0.84
cit patents Leiden-modularity	0	3	1.00	1.00	0.37
wiki topcats Leiden-modularity	0	100	1.00	1.00	0.74
wiki talk Leiden-modularity	0	6	1.00	1.00	0.03
orkut Leiden-modularity	0	86	1.00	1.00	0.91
cen Leiden-modularity	0	76	1.00	1.00	0.24
open citations Leiden-modularity	0	1	1.00	0.99	0.65
median values	0	23	1.00	1.00	0.65
cit hepph Leiden-0.5	0	0	0.85	0.16	0.16
cit patents Leiden-0.5	0	0	0.72	0.01	0.01
wiki topcats Leiden-0.5	0	0	0.68	0.02	0.02
wiki talk Leiden-0.5	0	0	0.04	0.00	0.00
orkut Leiden-0.5	0	0	0.89	0.30	0.30
cen Leiden-0.5	0	0	0.09	0.01	0.01
open citations Leiden-0.5	0	0	0.79	0.06	0.06
median values	0	0	0.72	0.06	0.06
cit hepph Leiden-0.1	0	5	0.94	0.63	0.63
cit patents Leiden-0.1	0	4	0.97	0.26	0.26
wiki topcats Leiden-0.1	0	6	0.90	0.28	0.27
wiki talk Leiden-0.1	0	24	0.10	0.00	0.00
orkut Leiden-0.1	0	4	0.95	0.65	0.64
cen Leiden-0.1	0	87	0.40	0.04	0.04
open citations Leiden-0.1	0	5	0.93	0.44	0.43
median values	0	6	0.93	0.28	0.27
cit hepph Leiden-0.01	0	44	0.96	0.91	0.85
cit patents Leiden-0.01	0	86	0.98	0.86	0.64
wiki topcats Leiden-0.01	0	75	0.95	0.84	0.65
wiki talk Leiden-0.01	0	52	0.34	0.03	0.01
orkut Leiden-0.01	0	83	0.96	0.94	0.86
cen Leiden-0.01	0	98	0.77	0.25	0.15
open citations Leiden-0.01	0	63	0.92	0.82	0.70
median values	0	75	0.95	0.84	0.65
cit hepph Leiden-0.001	0	19	0.97	0.94	0.85
cit patents Leiden-0.001	0	74	0.99	0.98	0.59
wiki topcats Leiden-0.001	0	40	0.95	0.92	0.66
wiki talk Leiden-0.001	0	60	0.80	0.26	0.02
orkut Leiden-0.001	0	55	0.97	0.96	0.84
cen Leiden-0.001	0	99	0.85	0.62	0.24
open citations Leiden-0.001	0	28	0.93	0.89	0.70
median values	0	55	0.95	0.92	0.66
cit hepph Leiden-0.0001	0	6	0.98	0.96	0.84
cit patents Leiden-0.0001	0	36	1.00	0.99	0.51
wiki topcats Leiden-0.0001	0	9	0.96	0.94	0.70
wiki talk Leiden-0.0001	0	61	0.98	0.76	0.02
orkut Leiden-0.0001	0	6	0.95	0.94	0.85
cen Leiden-0.0001	0	96	0.91	0.81	0.26
open citations Leiden-0.0001	0	7	0.94	0.92	0.68
median values	0	9	0.96	0.94	0.68

On the real-world datasets we examined, we see that (1) Clusterings using Leiden-CPM with high resolution values (0.1 or 0.5) generally had a large drop in node coverage (at least 30% other than for wiki_talk) after Stage 2 of the CM pipeline but very little drop (at most 1%) after Stage 3; this reflects the presence of small clusters but very few poorly connected clusters (0% for $r = 0.5$ and median of 6% for $r = 0.1$). (2) Clusterings produced using Leiden-CPM with small resolution values (at most 0.01) or Leiden-mod had very small drops in node coverage after Stage 2 but a large drop in node coverage after Stage 3 of the CM pipeline; this reflects the small number of small clusters and the large number of poorly connected clusters.

D Experiments on LFR Networks

D.1 Comparing LFR networks to their real-world networks

Figures C and K show a comparison between the degree distribution and community size distributions of the six real-world networks and their corresponding LFR networks for different Leiden clusterings. Tables K and L show various parameters of the real-world networks and their corresponding LFR graphs.

D.2 Impact of CM on clustering accuracy on LFR networks

For the accuracy experiments in the paper, we compared the accuracy of the pre-CM and post-CM clusterings of the LFR networks with respect to the LFR ground-truth communities. We used the normalized mutual information (NMI), adjusted mutual information (AMI), and adjusted Rand index (ARI) criteria that are commonly used for partition comparison, as implemented in the Scikit-learn library. The nodes in the original network that are removed from the post-CM clustering are added back as singletons (with distinct clusters), and hence the partitions are compared on the whole set of nodes.

Table K: Properties of the real-world networks and their LFR networks for Leiden-CPM clusterings. Parameters: r stands for resolution value used for generating the Leiden-CPM clustering, μ stands for the mixing parameter, and τ_1 and τ_2 are the *estimated* power-law exponents for the degree and the community size distributions respectively. c_{local} refers to the average local clustering coefficient and c_{global} refers to the global clustering coefficient.

network	r	nodes	edges	avg deg	μ	τ_1	τ_2	c_{local}	c_{global}
open_citations	0.0001	75,025,194	1,363,303,678	36.34	0.407	2.974	4.045	0.105	0.016
open_citations_LFR	0.0001	3,000,000	55,134,095	36.76	0.407	2.978	4.036	0.296	0.293
open_citations	0.001	75,025,194	1,363,303,678	36.34	0.500	2.974	4.375	0.105	0.016
open_citations_LFR	0.001	3,000,000	55,067,530	36.71	0.500	2.980	4.372	0.179	0.190
open_citations	0.01	75,025,194	1,363,303,678	36.34	0.602	2.974	5.205	0.105	0.016
open_citations_LFR	0.01	3,000,000	54,801,081	36.53	0.602	2.983	5.195	0.085	0.090
open_citations	0.1	75,025,194	1,363,303,678	36.34	0.711	2.974	6.194	0.105	0.016
open_citations_LFR	0.1	3,000,000	54,906,125	36.60	0.711	2.980	6.192	0.033	0.032
open_citations	0.5	75,025,194	1,363,303,678	36.34	0.871	2.974	6.152	0.105	0.016
open_citations_LFR	0.5	3,000,000	55,104,605	36.74	0.871	2.978	6.135	0.004	0.002
CEN	0.0001	13,989,436	92,051,051	13.16	0.402	2.618	2.259	0.138	0.002
CEN_LFR	0.0001	3,000,000	20,817,560	13.88	0.402	2.620	2.269	0.188	0.057
CEN	0.001	13,989,436	92,051,051	13.16	0.522	2.618	2.368	0.138	0.002
CEN_LFR	0.001	3,000,000	20,809,023	13.87	0.522	2.620	2.372	0.120	0.040
CEN	0.01	13,989,436	92,051,051	13.16	0.645	2.618	5.420	0.138	0.002
CEN_LFR	0.01	3,000,000	20,554,876	13.70	0.646	2.616	5.402	0.041	0.049
CEN	0.1	13,989,436	92,051,051	13.16	0.879	2.618	6.184	0.138	0.002
CEN_LFR	0.1	3,000,000	20,743,710	13.83	0.878	2.622	6.152	0.001	0.003
CEN	0.5	13,989,436	92,051,051	13.16	0.988	2.618	3.270	0.138	0.002
CEN_LFR	0.5	3,000,000	20,821,520	13.88	0.988	2.620	3.296	0.000	0.000
cit_patents	0.0001	3,774,768	16,518,947	8.75	0.211	4.017	2.981	0.092	0.067
cit_patents_LFR	0.0001	3,774,768	15,640,593	8.29	0.211	4.000	2.974	0.379	0.265
cit_patents	0.001	3,774,768	16,518,947	8.75	0.284	4.017	4.830	0.092	0.067
cit_patents_LFR	0.001	3,774,768	15,642,211	8.29	0.284	4.024	4.829	0.208	0.176
cit_patents	0.01	3,774,768	16,518,947	8.75	0.382	4.017	2.565	0.092	0.067
cit_patents_LFR	0.01	3,774,768	15,640,109	8.29	0.382	4.000	2.573	0.163	0.113
cit_patents	0.1	3,774,768	16,518,947	8.75	0.511	4.017	4.639	0.092	0.067
cit_patents_LFR	0.1	3,774,768	15,643,333	8.29	0.511	4.010	4.623	0.091	0.079
cit_patents	0.5	3,774,768	16,518,947	8.75	0.805	4.017	4.162	0.092	0.067
cit_patents_LFR	0.5	3,774,768	15,606,781	8.27	0.807	4.004	4.124	0.007	0.007
cit_hepph	0.0001	34,546	420,877	24.37	0.086	3.631	2.313	0.296	0.146
cit_hepph_LFR	0.0001	34,546	431,138	24.96	0.086	3.632	2.280	0.567	0.339
cit_hepph	0.001	34,546	420,877	24.37	0.219	3.631	1.465	0.296	0.146
cit_hepph_LFR	0.001	34,546	431,138	24.96	0.219	3.632	1.659	0.073	0.047
cit_hepph	0.01	34,546	420,877	24.37	0.384	3.631	1.825	0.296	0.146
cit_hepph_LFR	0.01	34,546	430,104	24.90	0.384	3.632	2.089	0.189	0.116
cit_hepph	0.1	34,546	420,877	24.37	0.570	3.631	2.333	0.296	0.146
cit_hepph_LFR	0.1	34,546	431,138	24.96	0.570	3.632	6.137	0.017	0.014
cit_hepph	0.5	34,546	420,877	24.37	0.781	3.631	2.790	0.296	0.146
cit_hepph_LFR	0.5	34,546	431,138	24.96	0.781	3.632	3.544	0.005	0.004
wiki_topcats	0.0001	1,791,489	25,444,207	28.41	0.379	2.430	1.645	0.276	0.002
wiki_topcats_LFR	0.0001	1,791,489	24,504,754	27.36	0.379	2.440	1.684	0.112	0.034
wiki_topcats	0.001	1,791,489	25,444,207	28.41	0.544	2.430	1.913	0.276	0.002
wiki_topcats_LFR	0.001	1,791,489	24,504,163	27.36	0.544	2.440	1.927	0.078	0.023
wiki_topcats	0.01	1,791,489	25,444,207	28.41	0.682	2.430	2.429	0.276	0.002
wiki_topcats_LFR	0.01	1,791,489	24,491,676	27.34	0.682	2.441	2.432	0.059	0.024
wiki_topcats	0.1	1,791,489	25,444,207	28.41	0.791	2.430	3.281	0.276	0.002
wiki_topcats_LFR	0.1	1,791,489	24,346,081	27.18	0.793	2.445	3.281	0.023	0.023
wiki_topcats	0.5	1,791,489	25,444,207	28.41	0.902	2.430	4.045	0.276	0.002

network	r	nodes	edges	average degree	μ	τ_1	τ_2	c_{local}	c_{global}
wiki_talk	0.0001	2,394,385	4,659,565	3.89	0.170	1.901	2.824	0.201	0.002
wiki_talk_LFR	0.0001	2,394,385	3,240,464	2.71	0.162	2.138	2.816	0.252	0.116
wiki_talk	0.001	2,394,385	4,659,565	3.89	0.346	1.901	1.901	0.201	0.002
wiki_talk_LFR	0.001	2,394,385	3,299,276	2.76	0.342	2.081	1.943	0.053	0.031
wiki_talk	0.01	2,394,385	4,659,565	3.89	0.754	1.901	1.917	0.201	0.002
wiki_talk_LFR	0.01	2,394,385	3,297,120	2.75	0.753	2.062	2.277	0.001	0.002
wiki_talk	0.1	2,394,385	4,659,565	3.89	0.941	1.901	2.228	0.201	0.002
wiki_talk_LFR	0.1	2,394,385	3,296,124	2.75	0.940	2.062	2.340	0.000	0.000
wiki_talk	0.5	2,394,385	4,659,565	3.89	0.984	1.901	3.064	0.201	0.002
orkut	0.0001	3,072,441	117,185,083	76.28	0.379	2.967	1.440	0.170	0.041
orkut	0.001	3,072,441	117,185,083	76.28	0.520	2.967	2.234	0.170	0.041
orkut	0.01	3,072,441	117,185,083	76.28	0.627	2.967	1.943	0.170	0.041
orkut	0.1	3,072,441	117,185,083	76.28	0.744	2.967	2.182	0.170	0.041
orkut	0.5	3,072,441	117,185,083	76.28	0.867	2.967	2.361	0.170	0.041

Table L: Properties of the real-world networks and their LFR networks for Leiden-modularity clusterings. Parameters: μ stands for the mixing parameter, and τ_1 and τ_2 are the *estimated* power-law exponents for the degree and the community size distributions respectively. c_{local} refers to the average local clustering coefficient and c_{global} refers to the global clustering coefficient.

network	nodes	edges	avg. deg.	μ	τ_1	τ_2	c_{local}	c_{global}
open_citations	75,025,194	1,363,303,678	36.34	0.129	2.974	2.697	0.105	0.016
open_citations_LFR	3,000,000	55,128,496	36.75	0.129	2.978	2.707	0.573	0.324
CEN	13,989,436	92,051,051	13.16	0.180	2.618	1.255	0.138	0.002
CEN_LFR	3,000,000	20,821,202	13.88	0.180	2.620	1.489	0.207	0.056
cit_patents	3,774,768	16,518,947	8.75	0.114	4.017	2.365	0.092	0.067
cit_patents_LFR	3,774,768	15,648,081	8.29	0.114	4.000	2.361	0.277	0.190
cit_hepph	34,546	420,877	24.37	0.155	3.631	1.305	0.296	0.146
cit_hepph_LFR	34,546	431,138	24.96	0.155	3.632	1.525	0.056	0.039
wiki_topcats	1,791,489	25,444,207	28.41	0.199	2.430	3.961	0.276	0.002
wiki_topcats_LFR	1,791,489	23,581,074	26.33	0.200	2.454	3.947	0.402	0.412
wiki_talk	2,394,385	4,659,565	3.89	0.115	1.901	2.074	0.201	0.002
wiki_talk_LFR	2,394,385	3,278,574	2.74	0.114	2.082	2.074	0.156	0.069
orkut	3,072,441	117,185,083	76.28	0.171	2.967	2.153	0.170	0.041

Table M: Percentage of LFR ground-truth clusters of size at least 11 that are disconnected. The LFR graphs are generated based on parameters derived from Leiden-modularity and Leiden-CPM (with five resolution values) clusterings for each of the six real-world networks. "N.A." means the statistic is unavailable because the LFR network for that condition could not be created.

		Leiden-CPM				
	Leiden-mod	0.0001	0.001	0.01	0.1	0.50
open_citations	0	0	0	0	0	0.001
CEN	0	0	0	0.04	90.81	100
cit_hepph	0	0	0	0	0	0
cit_patents	0	0	0	0	0.002	14.69
wiki_talk	70.72	66.36	87.03	100	100	N.A.
wiki_topcats	0	0	0	0	0.013	N.A.

D.3 Additional discussion for CM on LFR networks

D.3.1 The clustering coefficient

The LFR networks are constructed based on Leiden-modularity and Leiden-CPM clusterings with different values for the resolution parameter r. After the clustering is computed, the mixing parameter for the clustered real-world graph is computed. This mixing parameter value, along with the degree sequence and the cluster size sequence, is then given to LFR, which produces a synthetic network. The local clustering coefficient is calculated based on the synthetic network produced by LFR.

We report the Spearman rank correlation coefficient between the mixing parameter given as input to LFR and the resultant average local clustering coefficient. For the case of networks computed using Leiden-CPM clusterings, we also report the Spearman rank correlation coefficient between the resolution parameter r given to Leiden-CPM and the mixing parameter we give as input to LFR, and to the average local clustering coefficient. The implementation found at https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.spearmanr.html was used for the calculations.

- Spearman rank correlation coefficient between mixing parameter to average local clustering coefficient: -0.870
- Spearman rank correlation coefficient between resolution parameter to the mixing parameter: 0.839
- Spearman rank correlation coefficient between resolution parameter to average local clustering coefficient: -0.817

These strong negative correlation between the mixing parameter, which is given as input to LFR, and the local clustering coefficient of the resultant network suggests that the mixing parameter acts as effectively as a way of tuning the local clustering coefficients. The other two rank correlation coefficients we computed have to do with the impact of the resolution parameter r for the Leiden-CPM clusterings used as input for LFR. The Spearman rank correlation coefficient between the resolution parameter and the mixing parameter is strong and positive, showing that large resolution parameter values (e.g., 0.1 and 0.5) produce networks with high mixing parameter values, and conversely low resolution parameter values (e.g., 0.0001) produce networks with low mixing parameters. We also see, as an obvious consequence, that there is a strong negative correlation between the resolution parameter and the average local clustering coefficient. Thus, because our simulation study varied the Leiden-CPM clustering resolution parameter substantially, we obtain a wide range of local correlation coefficients, through the indirect effect on the mixing parameter.

As shown in Tables K and L, the average local clustering coefficients for the five corresponding real-world networks ranged from 0.092 to 0.296. In contrast, the clustering coefficients for the LFR networks computed using parameters derived from the different clusterings of these networks ranged from 0.000 to 0.573. Thus, the LFR networks have a larger range of average clustering coefficient values than the real-world networks, including ones that are much smaller and ones that are much larger.

In this experiment, we evaluated the impact of the clustering coefficient and the impact of CM on clustering accuracy on LFR networks, when used with Leiden-CPM at different resolution values. As seen in Figure B, in general the impact of CM on clustering accuracy is generally neutral when the local clustering coefficient is small (with a few outliers that show deleterious impact), but then becomes generally beneficial as the average local clustering coefficient increases. Since the local clustering coefficient reflects the "clusterability" of the network, this trend suggests that networks with sufficient clusterability will be better clustered by Leiden when followed CM.

A detailed look at each collection of LFR networks (one such collection per real-world network) is shown in Figure A. This figure shows that for some real-world networks, there is little impact of CM on clustering accuracy, but for others there is a clear trend of improvement for large average clustering coefficients.

Putting these results together, they suggest that for very low average clustering coefficients, below about 0.1, CM may be somewhat deleterious, but above this value it is either neutral or beneficial.

We also note that the range of average clustering coefficients on the seven real-world networks ranged from 0.092 to 0.296. Hence, the range in which CM has a tendency to be deleterious on LFR networks is generally below the observed range for these real-world networks. Moreover, small average local clustering coefficients indicate that the network is not very "clusterable". Hence, these trends suggest that on networks

that have sufficient clusterability, such as the real-world networks we explored, CM is likely to be neutral to beneficial.

D.3.2 Impact of each stage of the CM pipeline

In Table N, we show the percent connected, percent disconnected, and node coverage after Stages 2 and 3 in the CM pipeline. As promised, all Leiden clusters are connected, and so this focus is on the node coverage and how it changes in these two stages. Focusing on the nc2 results (i.e., node coverage by non-singleton clusters), we note that nearly all Leiden clusterings start with 100% node coverage. We see in general that Leiden-CPM responses to the stages in the CM pipeline depends on the resolution parameter value. For CPM-optimization with the largest resolution parameter values ($r = 0.5$), Stage 2 removes all the non-singleton clusters, and we confirm that these are due to all the clusters being below size 11. For $r = 0.1$, there is typically a large drop in node coverage in Stage 2, and then no drop (or a very small drop) in Stage 3; further investigation confirms that the drop in Stage 2 is entirely due to removing small clusters. For the two smallest resolution parameter values (i.e., $r = 0.001, 0.0001$) and Leiden-modularity, there is no drop in Stage 2, and Stage 3 has variable impact: sometimes no drop, but sometimes a substantial drop in node coverage. The intermediate resolution parameter ($r = 0.01$) shows a mixture of these trends, but is closer to the small resolution parameters.

Thus, Leiden clustering of LFR networks shows similar trends as on real-world networks, but with somewhat reduced impact from Stage 3, compared to real-world networks.

Table N: Connectivity and node coverage through different stages of the CM pipeline on LFR simulated datasets. "% disc." is the percentage of clusters of size at least two that are disconnected. "% poor" is the percentage of non-singleton clusters that are poorly connected (i.e., have an edge cut that is at most $\log_{10} n$). The terms nc11 and nc2 refer to the proportion of the network covered by clusters of size at least 11 or at least 2, respectively. N/A refers to clusterings in which all nodes were assigned to singleton clusters. "Stage 2" is the filtering step, which removes all small clusters and all tree clusters. "Stage 3" indicates that stage 3 (which finds and processes poorly-connected clusters) has completed but not also Stage 4, the post-filtering step in which small clusters are removed.

	% disc.	% poor	nc11	nc2
cit_hepph leiden mod	0	24	1.00	1.00
+ stage 2 (filter)	0	24	1.00	1.00
+ stage 3	0	0	0.99	0.99
cit_patents leiden mod	0	100	1.00	1.00
+ stage 2 (filter)	0	100	1.00	1.00
+ stage 3	0	0	0.63	0.63
wiki_topcats leiden mod	0	99	1.00	1.00
+ stage 2 (filter)	0	99	1.00	1.00
+ stage 3	0	0	0.98	0.98
cen leiden mod	0	100	1.00	1.00
+ stage 2 (filter)	0	100	1.00	1.00
+ stage 3	0	0	0.97	0.97
oc leiden mod	0	100	1.00	1.00
+ stage 2 (filter)	0	100	1.00	1.00
+ stage 3	0	0	0.99	0.99
cit_hepph leiden cpm 0.5	0	0	0.00	0.98
+ stage 2 (filter)	N/A	N/A	N/A	N/A
cit_patents leiden cpm 0.5	0	0	0.00	0.95
+ stage 2 (filter)	0	0	0.00	0.00
+ stage 3	0	0	0.00	0.00
cen leiden cpm 0.5	0	0	0.00	0.96
+ stage 2 (filter)	N/A	N/A	N/A	N/A
oc leiden cpm 0.5	0	0	0.00	0.99
+ stage 2 (filter)	0	0	0.00	0.00

+ stage 3	0	0	0.00	0.00
cit_hepph leiden cpm 0.1	0	1	0.33	1.00
+ stage 2 (filter)	0	2	0.33	0.33
+ stage 3	0	0	0.33	0.33
cit_patents leiden cpm 0.1	0	1	0.37	1.00
+ stage 2 (filter)	0	1	0.37	0.37
+ stage 3	0	0	0.37	0.37
wiki_topcats leiden cpm 0.1	0	0	0.29	1.00
+ stage 2 (filter)	0	1	0.29	0.29
+ stage 3	0	0	0.29	0.29
cen leiden cpm 0.1	0	0	0.00	1.00
+ stage 2 (filter)	0	20	0.00	0.00
+ stage 3	0	0	0.00	0.00
oc leiden cpm 0.1	0	0	0.93	1.00
+ stage 2 (filter)	0	0	0.93	0.93
+ stage 3	0	0	0.93	0.93
cit_hepph leiden cpm 0.01	0	0	1.00	1.00
+ stage 2 (filter)	0	0	1.00	1.00
+ stage 3	0	0	1.00	1.00
cit_patents leiden cpm 0.01	0	62	0.98	1.00
+ stage 2 (filter)	0	61	0.98	0.98
+ stage 3	0	0	0.58	0.58
wiki_topcats leiden cpm 0.01	0	72	0.99	1.00
+stage 2 (filter)	0	72	0.99	0.99
+ stage 3	0	0	0.64	0.64
cen leiden cpm 0.01	0	23	1.00	1.00
+ stage 2 (filter)	0	23	1.00	1.00
+ stage 3	0	0	0.99	0.99
oc leiden cpm 0.01	0	0	1.00	1.00
+ stage 2 (filter)	0	0	1.00	1.00
+ stage 3	0	0	1.00	1.00
cit_hepph leiden cpm 0.001	0	27	1.00	1.00
+ stage 2 (filter)	0	27	1.00	1.00
+ stage 3	0	0	1.00	1.00
cit_patents leiden cpm 0.001	0	95	1.00	1.00
+ stage 2 (filter)	0	95	1.00	1.00
+ stage 3	0	0	1.00	1.00
wiki_topcats leiden cpm 0.001	0	71	1.00	1.00
+stage 2 (filter)	0	71	1.00	1.00
+ stage 3	0	0	0.92	0.92
cen leiden cpm 0.001	0	84	1.00	1.00
+ stage 2 (filter)	0	84	1.00	1.00
+ stage 3	0	0	0.70	0.70
oc leiden cpm 0.001	0	47	1.00	1.00
+ stage 2 (filter)	0	47	1.00	1.00
+ stage 3	0	0	1.00	1.00
cit_hepph leiden cpm 0.0001	0	67	1.00	1.00
+ stage 2 (filter)	0	67	1.00	1.00
+ stage 3	0	0	1.00	1.00
cit_patents leiden cpm 0.0001	0	100	1.00	1.00
+ stage 2 (filter)	0	100	1.00	1.00
+ stage 3	0	0	0.56	0.56

wiki_topcats leiden cpm 0.0001	0	83	1.00	1.00
+stage 2 (filter)	0	83	1.00	1.00
+ stage 3	0	0	0.98	0.98
cen leiden cpm 0.0001	0	100	1.00	1.00
+ stage 2 (filter)	0	100	1.00	1.00
+ stage 3	0	0	0.85	0.85
oc leiden cpm 0.0001	0	100	1.00	1.00
+ stage 2 (filter)	0	100	1.00	1.00
+ stage 3	0	0	1.00	1.00

Table O: Impact of CM on LFR networks with many disconnected ground-truth clusters, combined The impact of CM on Leiden clustering accuracy is shown for each of the LFR networks that had a large fraction of disconnected clusters (specifically, at least 50% of the clusters of size at least 11 are disconnected). For a given accuracy metric (NMI, AMI, or ARI), Δ refers to the difference in the accuracy measure for pre-CM and post-CM clustering, so that a positive value indicates that CM improves accuracy and a negative value indicates that CM decreases accuracy. Results shown here indicate that CM is neutral to beneficial for NMI but reduces accuracy for the other two metrics on these networks, each of which has a high percentage of disconnected clusters.

	Clust. Coeff.	Mix. Param.	Δ NMI	Δ AMI	Δ ARI	% disconn. $n \geq 11$	% well-conn. $n \geq 11$
wiki_talk modularity	0.156	0.114	0.037	-0.500	-0.234	75	0
wiki_talk 0.0001	0.252	0.162	0.019	-0.427	-0.240	65	0
wiki_talk 0.001	0.053	0.342	0.029	-0.327	-0.137	89	0
wiki_talk 0.01	0.001	0.753	0.090	-0.105	-0.013	100	0
wiki_talk 0.1	0.000	0.940	0.070	-0.018	-0.001	100	0
CEN 0.1	0.001	0.878	0.050	-0.068	-0.024	91	0
CEN 0.5	0.000	0.988	0.093	-0.003	-0.000	100	0

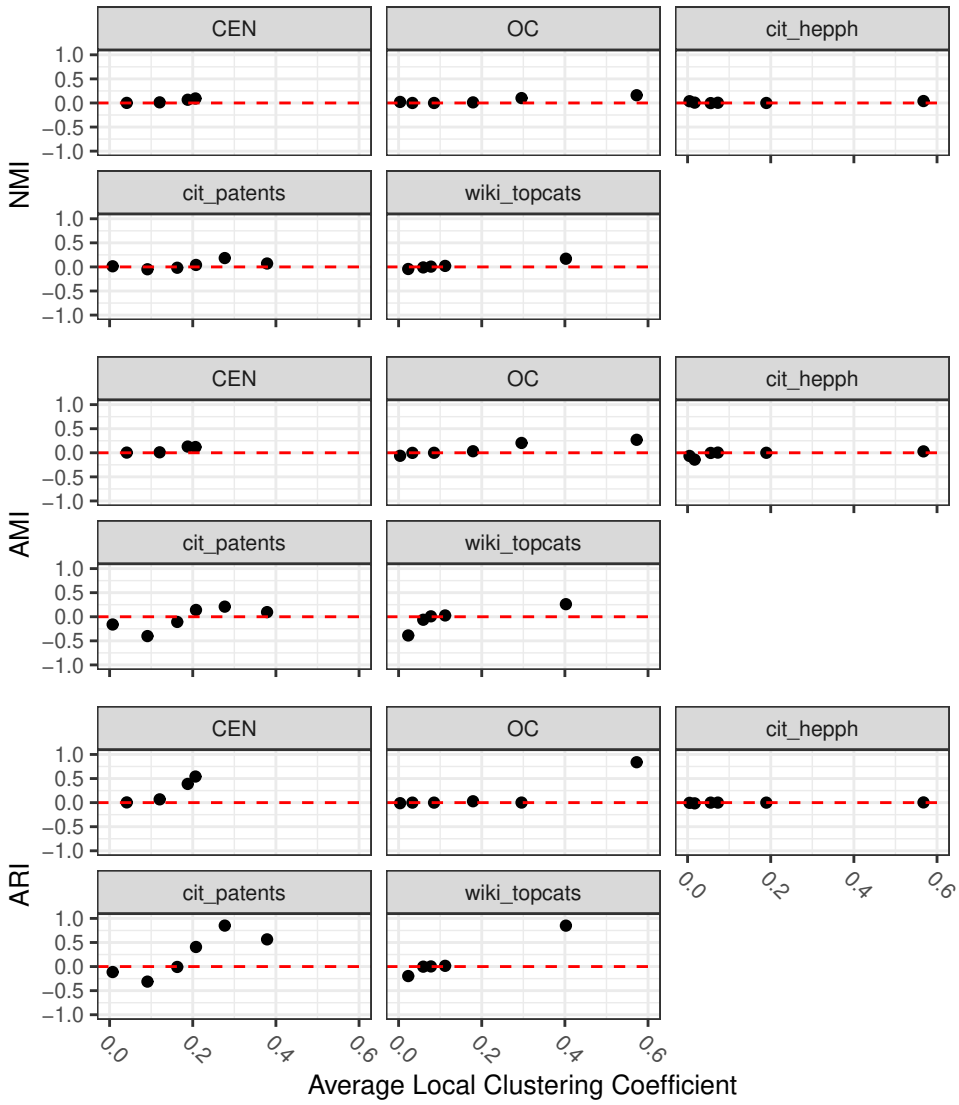

Figure A: Impact of CM on accuracy on LFR networks with respect to average local clustering coefficient, by network The y-axis is the change in the clustering accuracy produced by using CM, with positive values indicating that CM improves accuracy for the selected criterion (NMI, AMI, or ARI), and negative values indicating that CM reduced accuracy. The average local clustering coefficients for the individual real-world networks: CEN: 0.138, OC: 0.105, cit_hepph: 0.296, cit_pat: 0.092, and wiki_topcat: 0.276.

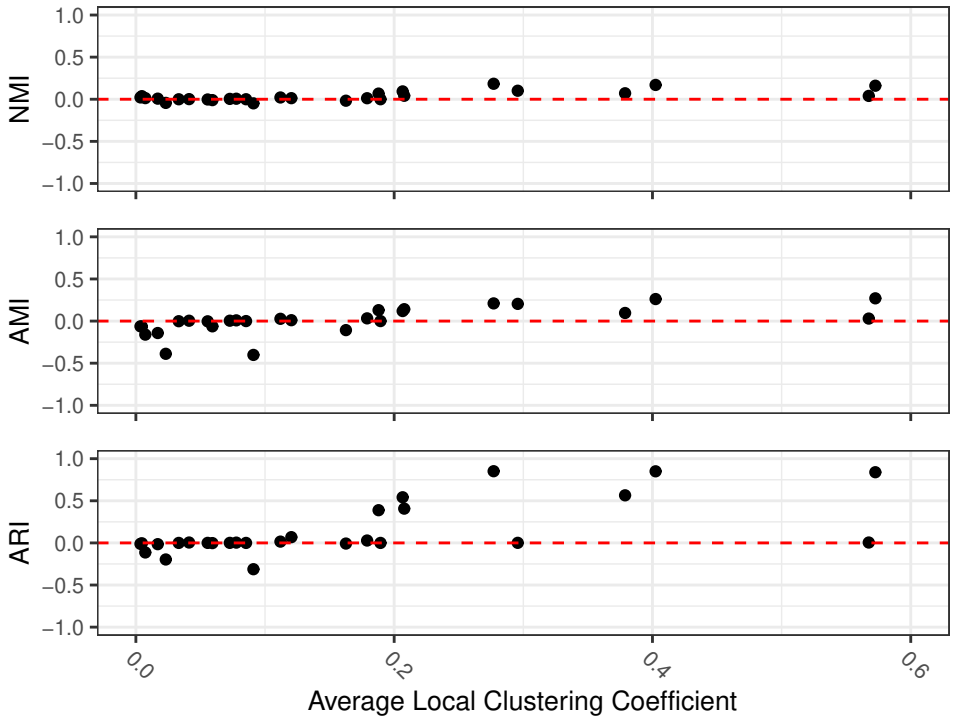

Figure B: Impact of CM on Leiden clusterings on LFR networks, as a function of the average local clustering coefficient, combined. The y-axis is the change in the clustering accuracy produced by using CM, with positive values indicating that CM improves accuracy for the selected criterion (NMI, AMI, or ARI), and negative values indicating that CM reduced accuracy. The average local clustering coefficient values for the corresponding real-world networks range from 0.092 to 0.276.

E Experiments on Erdős-Rényi Graphs

We generated a collection of Erdős-Rényi graphs, each with 10,000 nodes but varying values for p (the probability for each pair of vertices to be adjacent), with 10 replicates for each model condition specified by p. Using default settings for CM, we performed a series of experiments. For each model condition, we computed Leiden-CPM+CM clusterings using different resolution parameter values, ranging from 0.0001 to 0.5; we also generated Leiden-modularity+CM clusterings. We report the average (across the 10 replicates) of the number of clusters both pre-CM and post-CM for each combination of method and model condition. We also report the number of nodes in clusters of size at least 2 and of size at least 11 for both pre-CM and post-CM combination. Finally, we explore the impact of changing the requirement for well-connectedness in CM for the model condition corresponding to the largest value of p (i.e., the densest Erdős-Rényi graphs).

Table P: Properties of the Erdős-Rényi graphs Empirical statistics of Erdős-Rényi graphs with 10,000 nodes and varying p. For each model condition (specified by the value p), we also show the average degree of the nodes (avg. deg.), number of isolated nodes (num. isol. nodes), average of the local clustering coefficients (local clus. coeff.), and global clustering coefficient (global clus. coeff.). The results in this table are averaged over 10 replicates.

p	0.00001	0.00005	0.0001	0.0005	0.001	0.005
avg. deg.	0.1	0.5	1.0	5.0	10.0	49.9
num. isol. nodes	9030	6074	3667	70	0	0
local clus. coeff.	0.0000	0.0002	0.0002	0.0005	0.0010	0.0050
global clus. coeff.	0.0000	0.0002	0.0002	0.0005	0.0010	0.0050

Table Q: Impact of CM on the Node Coverage in Erdős-Rényi graphs on clusters of size at least 2 The Erdős-Rényi graphs have $10,000$ nodes and varying p. We show average node coverage in the pre-CM and post-CM clusterings of the Erdős-Rényi graphs using Leiden-mod, IKC, and Leiden-CPM (resolution value shown parenthetically). We also show the average degree of the nodes (avg. deg.), number of isolated nodes (num. isol. nodes), average of the local clustering coefficients (local clus. coeff.), and global clustering coefficient (global clus. coeff.). The results in this table are averaged over 10 replicates.

p	0.00001	0.00005	0.0001	0.0005	0.001	0.005
avg. deg.	0.1	0.5	1.0	5.0	10.0	49.9
num. isol. nodes	9030	6074	3667	70	0	0
local clus. coeff.	0.0000	0.0002	0.0002	0.0005	0.0010	0.0050
global clus. coeff.	0.0000	0.0002	0.0002	0.0005	0.0010	0.0050
Leiden-Mod	0	90	2506	9927	10000	10000
Leiden-Mod-cm	0	0	7	3553	3760	9999
IKC	0	0	22	8538	8754	9748
IKC-cm	0	0	17	975	7554	9569
Leiden-CPM(0.5)	0	0	0	0	0	0
Leiden-CPM(0.5)-cm	0	0	0	0	0	0
Leiden-CPM(0.1)	0	0	0	0	1	559
Leiden-CPM(0.1)-cm	0	0	0	0	1	550
Leiden-CPM(0.01)	0	90	2322	9874	9969	10000
Leiden-CPM(0.01)-cm	0	0	0	2272	6690	9978
Leiden-CPM(0.001)	0	90	2506	9927	9998	10000
Leiden-CPM(0.001)-cm	0	0	4	6393	5114	10000
Leiden-CPM(0.0001)	0	90	2506	9901	10000	10000
Leiden-CPM(0.0001)-cm	0	0	7	964	9897	10000

Table R: Impact of CM on the Number of Clusters in Erdős-Rényi Graphs on clusters of size at least 11 Erdős-Rényi graphs with 10,000 nodes and different settings for p were clustered using Leiden-mod-cm, Leiden-CPM-cm (with varying resolution values), and IKC(k=2)-cm; we show the pre- and post-CM number of clusters for each method (here, post-CM means all four stages have completed). For each model condition, defined by p, we show the average degree of the nodes (avg. deg.), number of isolated nodes (num. isol. nodes), average local clustering coefficients (local clus. coeff.), and average global clustering coefficient (global clus. coeff). The results in this table are averaged over 10 replicates.

p	0.00001	0.00005	0.0001	0.0005	0.001	0.005
avg. deg.	0.1	0.5	1.0	5.0	10.0	49.9
num. isol. nodes	9030	6074	3667	70	0	0
local clus. coeff.	0.0000	0.0002	0.0002	0.0005	0.0010	0.0050
global clus. coeff.	0.0000	0.0002	0.0002	0.0005	0.0010	0.0050
Leiden-Mod	0	7	86	40	22	8
Leiden-Mod-cm	0	0	0	41	22	8
IKC(2)	0	0	0	1	2	2
IKC(2)-cm	0	0	0	1	2	2
Leiden-CPM(0.5)	0	0	0	0	0	0
Leiden-CPM(0.5)-cm	0	0	0	0	0	0
Leiden-CPM(0.1)	0	0	0	0	0	48
Leiden-CPM(0.1)-cm	0	0	0	0	0	47
Leiden-CPM(0.01)	0	7	162	573	436	95
Leiden-CPM(0.01)-cm	0	0	0	155	362	95
Leiden-CPM(0.001)	0	7	96	96	24	1
Leiden-CPM(0.001)-cm	0	0	0	96	24	1
Leiden-CPM(0.0001)	0	7	80	1	1	1
Leiden-CPM(0.0001)-cm	0	0	0	1	1	1

Table S: Impact of increasing the threshold for well-connectedness in CM, when running Leiden-CPM(0.01)-CM on Erdős-Rényi (p=0.005) The first column specifies the threshold for well-connectedness. Num Nodes refers to the number of nodes in non-singleton clusters, after the CM pipeline completes. Num Clusters, Cluster Size Dist, and Mincut Size Dist refer to the clusters after the CM-pipeline completes; note that this does not include any clusters of size below 11. Results are averaged over 10 replicates.

Threshold	Num Nodes	Num Clusters	Cluster Size Dist	Mincut Size Dist
1 log 10	9978	95	17.1/105.1/315.8	1.0/1.3/1.8
2 log 10	4897	68	12.6/71.7/306.6	2.0/2.1/2.7
4 log 10	0	0	N/A	N/A

F Experiments on nPSO Networks

We use the implementation provided at https://github.com/biomedical-cybernetics/nPSO_model/blob/master/nPSO_model.m and run the following command to invoke nPSO:

```
distr = create_mixture_gaussian_gamma_pdf(100);
[<adjacency matrix output>, <coordinates>, <clustering>, <distance matrix>] =
    nPSO_model(10000, <half of average degree>, <temperature>, 3, distr, 0)
```

Table T: Impact of CM on nPSO networks The results here are for the ten nPSO model conditions where the networks had at most 15% disconnected ground truth clusters. "deg" refers to average degree (i.e., it is 2M, where M is one of the required input parameters); nc2 refers to node coverage from non-singleton clusters. N/A refers to clusterings in which all nodes were assigned to singleton clusters. The notation "+ stage 2 (filter)" indicates that the pipeline is run through Stage 2, which filters small clusters and tree clusters, but does not go further. The notation "+ stage 3" indicates that the pipeline is run through Stage 3, but that Stage 4, the post-filtering step in which small clusters are removed, has not been run.

	deg.	temp.	% poor	nc2	nc11	NMI	ARI	AMI
leiden mod	8	0	18	1.00	1.00	0.92	0.63	0.91
+ stage 2 (filter)	8	0	18	1.00	1.00	0.92	0.63	0.91
+ stage 3	8	0	0	1.00	1.00	0.92	0.63	0.91
leiden cpm 0.5	8	0	0	0.93	0.00	0.74	0.07	0.47
+ stage 2 (filter)	8	0	N/A	N/A	N/A	N/A	N/A	N/A
+ stage 3	8	0	N/A	N/A	N/A	N/A	N/A	N/A
leiden cpm 0.1	8	0	1	1.00	0.79	0.82	0.25	0.72
+ stage 2 (filter)	8	0	1	0.79	0.79	0.79	0.22	0.64
+ stage 3	8	0	0	0.78	0.78	0.79	0.22	0.64
leiden cpm 0.01	8	0	3	1.00	1.00	0.95	0.82	0.94
+ stage 2 (filter)	8	0	3	1.00	1.00	0.95	0.82	0.94
+ stage 3	8	0	0	1.00	1.00	0.95	0.82	0.94
leiden cpm 0.001	8	0	18	1.00	1.00	0.92	0.66	0.92
+ stage 2 (filter)	8	0	18	1.00	1.00	0.92	0.66	0.92
+ stage 3	8	0	0	1.00	1.00	0.92	0.66	0.92
leiden cpm 0.0001	8	0	18	1.00	1.00	0.76	0.26	0.75
+ stage 2 (filter)	8	0	18	1.00	1.00	0.76	0.26	0.75
+ stage 3	8	0	0	1.00	1.00	0.76	0.26	0.75
leiden mod	8	0.3	14	1.00	1.00	0.88	0.51	0.87
+ stage 2 (filter)	8	0.3	14	1.00	1.00	0.88	0.51	0.87
+ stage 3	8	0.3	0	1.00	1.00	0.88	0.51	0.87
leiden cpm 0.5	8	0.3	0	0.89	0.00	0.72	0.04	0.37
+ stage 2 (filter)	8	0.3	N/A	N/A	N/A	N/A	N/A	N/A
+ stage 3	8	0.3	N/A	N/A	N/A	N/A	N/A	N/A
leiden cpm 0.1	8	0.3	3	1.00	0.78	0.82	0.26	0.72
+ stage 2 (filter)	8	0.3	3	0.78	0.78	0.79	0.24	0.65
+ stage 3	8	0.3	0	0.78	0.77	0.79	0.24	0.65
leiden cpm 0.01	8	0.3	12	1.00	1.00	0.97	0.93	0.97
+ stage 2 (filter)	8	0.3	13	1.00	1.00	0.97	0.93	0.97
+ stage 3	8	0.3	0	1.00	1.00	0.97	0.93	0.97
leiden cpm 0.001	8	0.3	18	1.00	1.00	0.89	0.55	0.89
+ stage 2 (filter)	8	0.3	18	1.00	1.00	0.89	0.55	0.89
+ stage 3	8	0.3	0	1.00	1.00	0.89	0.55	0.89
leiden cpm 0.0001	8	0.3	29	1.00	1.00	0.71	0.20	0.71
+ stage 2 (filter)	8	0.3	29	1.00	1.00	0.71	0.20	0.71
+ stage 3	8	0.3	0	1.00	1.00	0.71	0.20	0.71

leiden mod	8	0.6	86	1.00	1.00	0.80	0.34	0.79
+ stage 2 (filter)	8	0.6	86	1.00	1.00	0.80	0.34	0.79
+ stage 3	8	0.6	0	0.99	0.99	0.80	0.34	0.79
leiden cpm 0.5	8	0.6	0	0.88	0.00	0.71	0.03	0.30
+ stage 2 (filter)	8	0.6	N/A	N/A	N/A	N/A	N/A	N/A
+ stage 3	8	0.6	N/A	N/A	N/A	N/A	N/A	N/A
leiden cpm 0.1	8	0.6	5	1.00	0.57	0.79	0.20	0.66
+ stage 2 (filter)	8	0.6	9	0.57	0.57	0.75	0.16	0.50
+ stage 3	8	0.6	0	0.57	0.55	0.75	0.16	0.50
leiden cpm 0.01	8	0.6	56	1.00	1.00	0.98	0.96	0.98
+ stage 2 (filter)	8	0.6	58	1.00	1.00	0.98	0.96	0.98
+ stage 3	8	0.6	0	0.99	0.99	0.98	0.96	0.98
leiden cpm 0.001	8	0.6	93	1.00	1.00	0.83	0.41	0.83
+ stage 2 (filter)	8	0.6	93	1.00	1.00	0.83	0.41	0.83
+ stage 3	8	0.6	0	0.99	0.99	0.83	0.41	0.83
leiden cpm 0.0001	8	0.6	100	1.00	1.00	0.55	0.09	0.55
+ stage 2 (filter)	8	0.6	100	1.00	1.00	0.55	0.09	0.55
+ stage 3	8	0.6	0	0.98	0.98	0.55	0.09	0.55
leiden mod	16	0	0	1.00	1.00	0.88	0.50	0.87
+ stage 2 (filter)	16	0	0	1.00	1.00	0.88	0.50	0.87
+ stage 3	16	0	0	1.00	1.00	0.88	0.50	0.87
leiden cpm 0.5	16	0	0	0.98	0.36	0.78	0.14	0.61
+ stage 2 (filter)	16	0	0	0.36	0.36	0.71	0.08	0.32
+ stage 3	16	0	0	0.36	0.36	0.71	0.08	0.32
leiden cpm 0.1	16	0	0	1.00	0.98	0.88	0.46	0.83
+ stage 2 (filter)	16	0	0	0.98	0.98	0.87	0.46	0.82
+ stage 3	16	0	0	0.98	0.98	0.87	0.46	0.82
leiden cpm 0.01	16	0	0	1.00	1.00	0.98	0.93	0.98
+ stage 2 (filter)	16	0	0	1.00	1.00	0.98	0.93	0.98
+ stage 3	16	0	0	1.00	1.00	0.98	0.93	0.98
leiden cpm 0.001	16	0	0	1.00	1.00	0.84	0.41	0.83
+ stage 2 (filter)	16	0	0	1.00	1.00	0.84	0.41	0.83
+ stage 3	16	0	0	1.00	1.00	0.84	0.41	0.83
leiden cpm 0.0001	16	0	0	1.00	1.00	0.61	0.12	0.61
+ stage 2 (filter)	16	0	0	1.00	1.00	0.61	0.12	0.61
+ stage 3	16	0	0	1.00	1.00	0.61	0.12	0.61
leiden mod	16	0.3	0	1.00	1.00	0.83	0.40	0.82
+ stage 2 (filter)	16	0.3	0	1.00	1.00	0.83	0.40	0.82
+ stage 3	16	0.3	0	1.00	1.00	0.83	0.40	0.82
leiden cpm 0.5	16	0.3	0	0.95	0.05	0.75	0.08	0.49
+ stage 2 (filter)	16	0.3	0	0.05	0.05	0.67	0.01	0.05
+ stage 3	16	0.3	0	0.05	0.05	0.67	0.01	0.05
leiden cpm 0.1	16	0.3	1	1.00	0.96	0.89	0.54	0.85
+ stage 2 (filter)r	16	0.3	1	0.96	0.96	0.88	0.53	0.84
+ stage 3	16	0.3	0	0.96	0.96	0.88	0.53	0.84
leiden cpm 0.01	16	0.3	1	1.00	1.00	0.97	0.88	0.97
+ stage 2 (filter)	16	0.3	1	1.00	1.00	0.97	0.88	0.97
+ stage 3	16	0.3	0	1.00	1.00	0.97	0.88	0.97
leiden cpm 0.001	16	0.3	0	1.00	1.00	0.79	0.32	0.79
+ stage 2 (filter)	16	0.3	0	1.00	1.00	0.79	0.32	0.79
+ stage 3	16	0.3	0	1.00	1.00	0.79	0.32	0.79
leiden cpm 0.0001	16	0.3	20	1.00	1.00	0.51	0.07	0.51

+ stage 2 (filter)	16	0.3	20	1.00	1.00	0.51	0.07	0.51
+ stage 3	16	0.3	0	1.00	1.00	0.51	0.07	0.51
leiden mod	16	0.6	12	1.00	1.00	0.76	0.26	0.75
+ stage 2 (filter)	16	0.6	12	1.00	1.00	0.76	0.26	0.75
+ stage 3	16	0.6	0	1.00	1.00	0.76	0.26	0.75
leiden cpm 0.5	16	0.6	0	0.94	0.00	0.72	0.05	0.39
+ stage 2 (filter)	16	0.6	N/A	N/A	N/A	N/A	N/A	N/A
+ stage 3	16	0.6	N/A	N/A	N/A	N/A	N/A	N/A
leiden cpm 0.1	16	0.6	4	0.99	0.89	0.87	0.46	0.81
+ stage 2 (filter)	16	0.6	4	0.89	0.89	0.86	0.46	0.79
+ stage 3	16	0.6	0	0.89	0.89	0.86	0.46	0.79
leiden cpm 0.01	16	0.6	5	1.00	1.00	0.96	0.80	0.95
+ stage 2 (filter)	16	0.6	5	1.00	1.00	0.96	0.80	0.95
+ stage 3	16	0.6	0	1.00	1.00	0.96	0.80	0.95
leiden cpm 0.001	16	0.6	8	1.00	1.00	0.69	0.19	0.69
+ stage 2 (filter)	16	0.6	8	1.00	1.00	0.69	0.19	0.69
+ stage 3	16	0.6	0	1.00	1.00	0.69	0.19	0.69
leiden cpm 0.0001	16	0.6	0	1.00	1.00	0.26	0.02	0.26
+ filter	16	0.6	0	1.00	1.00	0.26	0.02	0.26
+ stage 3	16	0.6	0	1.00	1.00	0.26	0.02	0.26
leiden mod	32	0	0	1.00	1.00	0.82	0.36	0.81
+ stage 2 (filter)	32	0	0	1.00	1.00	0.82	0.36	0.81
+ stage 3	32	0	0	1.00	1.00	0.82	0.36	0.81
leiden cpm 0.5	32	0	0	0.99	0.79	0.82	0.27	0.72
+ stage 2 (filter)	32	0	0	0.79	0.79	0.80	0.25	0.66
+ stage 3	32	0	0	0.79	0.79	0.80	0.25	0.66
leiden cpm 0.1	32	0	0	1.00	0.99	0.96	0.86	0.95
+ stage 2 (filter)	32	0	0	0.99	0.99	0.96	0.86	0.95
+ stage 3	32	0	0	0.99	0.99	0.96	0.86	0.95
leiden cpm 0.01	32	0	0	1.00	1.00	0.92	0.65	0.92
+ stage 2 (filter)	32	0	0	1.00	1.00	0.92	0.65	0.92
+ stage 3	32	0	0	1.00	1.00	0.92	0.65	0.92
leiden cpm 0.001	32	0	0	1.00	1.00	0.71	0.20	0.71
+ stage 2 (filter)	32	0	0	1.00	1.00	0.71	0.20	0.71
+ stage 3	32	0	0	1.00	1.00	0.71	0.20	0.71
leiden cpm 0.0001	32	0	0	1.00	1.00	0.46	0.06	0.46
+ stage 2 (filter)	32	0	0	1.00	1.00	0.46	0.06	0.46
+ stage 3	32	0	0	1.00	1.00	0.46	0.06	0.46
leiden mod	32	0.3	0	1.00	1.00	0.79	0.32	0.78
+ stage 2 (filter)	32	0.3	0	1.00	1.00	0.79	0.32	0.78
+ stage 3	32	0.3	0	1.00	1.00	0.79	0.32	0.78
leiden cpm 0.5	32	0.3	0	0.98	0.44	0.78	0.17	0.61
+ stage 2 (filter)	32	0.3	0	0.44	0.44	0.73	0.13	0.40
+ stage 3	32	0.3	0	0.44	0.44	0.73	0.13	0.40
leiden cpm 0.1	32	0.3	0	1.00	0.98	0.98	0.94	0.97
+ stage 2 (filter)	32	0.3	0	0.98	0.98	0.97	0.94	0.97
+ stage 3	32	0.3	0	0.98	0.98	0.97	0.94	0.97
leiden cpm 0.01	32	0.3	0	1.00	1.00	0.89	0.57	0.89
+ stage 2 (filter)	32	0.3	0	1.00	1.00	0.89	0.57	0.89
+ stage 3	32	0.3	0	1.00	1.00	0.89	0.57	0.89
leiden cpm 0.001	32	0.3	0	1.00	1.00	0.68	0.17	0.67
+ stage 2 (filter)	32	0.3	0	1.00	1.00	0.68	0.17	0.67

+ stage 3	32	0.3	0	1.00	1.00	0.68	0.17	0.67
leiden cpm 0.0001	32	0.3	0	1.00	1.00	0.38	0.04	0.38
+ stage 2 (filter)	32	0.3	0	1.00	1.00	0.38	0.04	0.38
+ stage 3	32	0.3	0	1.00	1.00	0.38	0.04	0.38
leiden mod	32	0.6	0	1.00	1.00	0.72	0.22	0.72
+ stage 2 (filter)	32	0.6	0	1.00	1.00	0.72	0.22	0.72
+ stage 3	32	0.6	0	1.00	1.00	0.72	0.22	0.72
leiden cpm 0.5	32	0.6	0	0.97	0.08	0.75	0.09	0.49
+ stage 2 (filter)	32	0.6	0	0.08	0.08	0.68	0.02	0.08
+ stage 3	32	0.6	0	0.08	0.08	0.68	0.02	0.08
leiden cpm 0.1	32	0.6	0	1.00	0.96	0.96	0.87	0.95
+ filter	32	0.6	0	0.96	0.96	0.95	0.87	0.94
+ stage 3	32	0.6	0	0.96	0.96	0.95	0.87	0.94
leiden cpm 0.01	32	0.6	0	1.00	1.00	0.85	0.46	0.85
+ filter	32	0.6	0	1.00	1.00	0.85	0.46	0.85
+ stage 3	32	0.6	0	1.00	1.00	0.85	0.46	0.85
leiden cpm 0.001	32	0.6	0	1.00	1.00	0.51	0.08	0.51
+ stage 2 (filter)	32	0.6	0	1.00	1.00	0.51	0.08	0.51
+ stage 3	32	0.6	0	1.00	1.00	0.51	0.08	0.51
leiden cpm 0.0001	32	0.6	0	1.00	1.00	0.00	0.00	0.00
+ stage 2 (filter)	32	0.6	0	1.00	1.00	0.00	0.00	0.00
+ stage 3	32	0.6	0	1.00	1.00	0.00	0.00	0.00
leiden mod	32	0.9	22	1.00	1.00	0.64	0.14	0.63
+ stage 2 (filter)	32	0.9	22	1.00	1.00	0.64	0.14	0.63
+ stage 3	32	0.9	0	1.00	1.00	0.64	0.14	0.63
leiden cpm 0.5	32	0.9	0	0.97	0.02	0.71	0.06	0.38
+ stage 2 (filter)	32	0.9	0	0.02	0.02	0.67	0.00	0.02
+ stage 3	32	0.9	0	0.02	0.02	0.67	0.00	0.02
leiden cpm 0.1	32	0.9	3	1.00	0.89	0.89	0.65	0.85
+ stage 2 (filter)	32	0.9	3	0.89	0.89	0.89	0.65	0.84
+ stage 3	32	0.9	0	0.89	0.89	0.89	0.65	0.84
leiden cpm 0.01	32	0.9	6	1.00	1.00	0.84	0.44	0.84
+ stage 2 (filter)	32	0.9	6	1.00	1.00	0.84	0.44	0.84
+ stage 3	32	0.9	0	1.00	1.00	0.84	0.44	0.84
leiden cpm 0.001	32	0.9	67	1.00	1.00	0.35	0.03	0.35
+ stage 2 (filter)	32	0.9	67	1.00	1.00	0.35	0.03	0.35
+ stage 3	32	0.9	0	0.99	0.99	0.35	0.03	0.35
leiden cpm 0.0001	32	0.9	0	1.00	1.00	0.00	0.00	0.00
+ stage 2 (filter)	32	0.9	0	1.00	1.00	0.00	0.00	0.00
+ stage 3	32	0.9	0	1.00	1.00	0.00	0.00	0.00

G Additional Figures

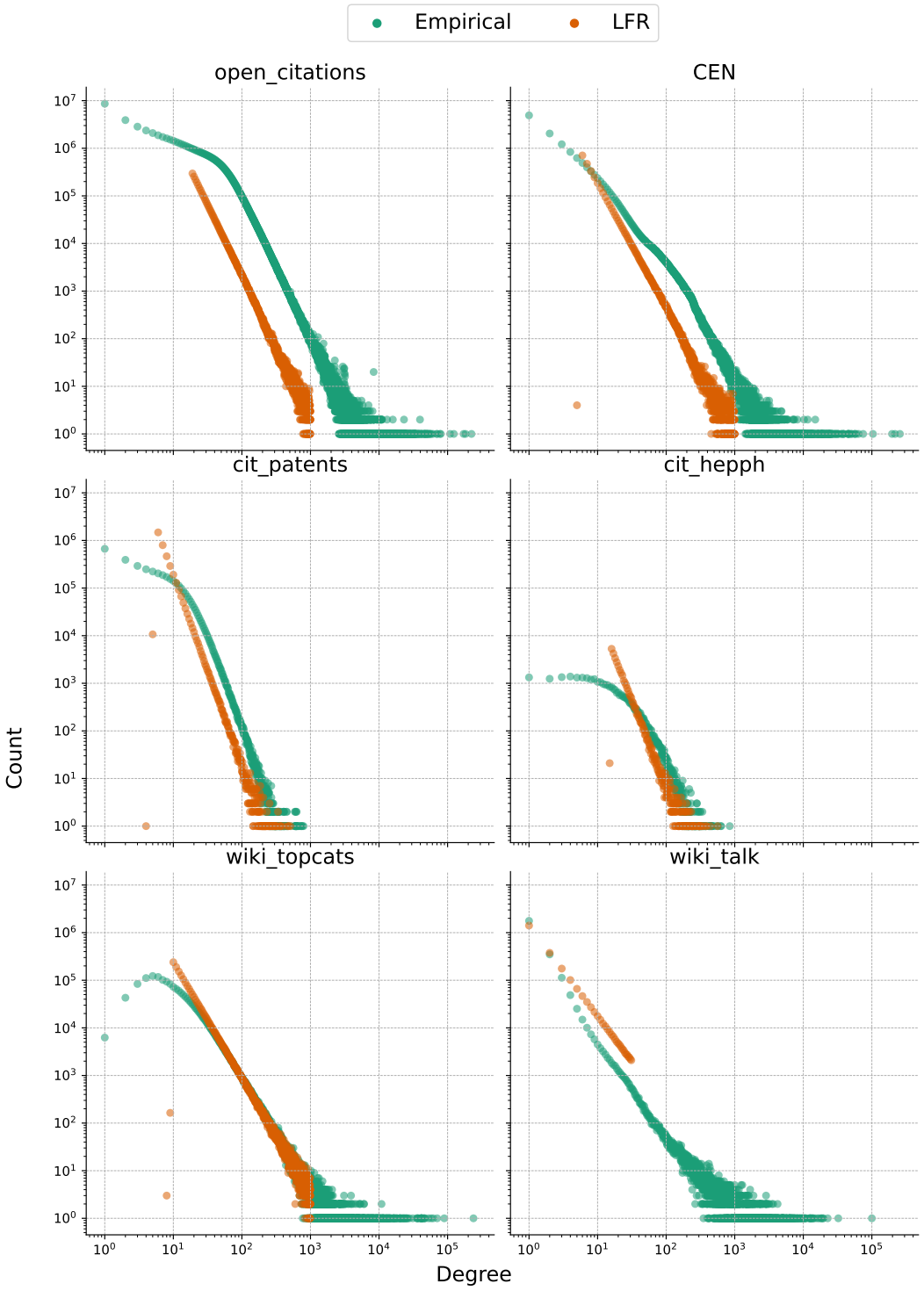

Figure C: Comparison between the degree distribution of real-world and LFR networks using CPM with $r = 0.01$. The LFR networks are produced to emulate the characteristics of their corresponding real-world network. For the CEN and the Open Citations networks, the number of nodes of the LFR network is 3,000,000, and for the other networks it exactly matches the number of nodes in its corresponding real-world network. The clustering method for the real-world networks was CPM with resolution parameter 0.01. The axes are shown in log scale.

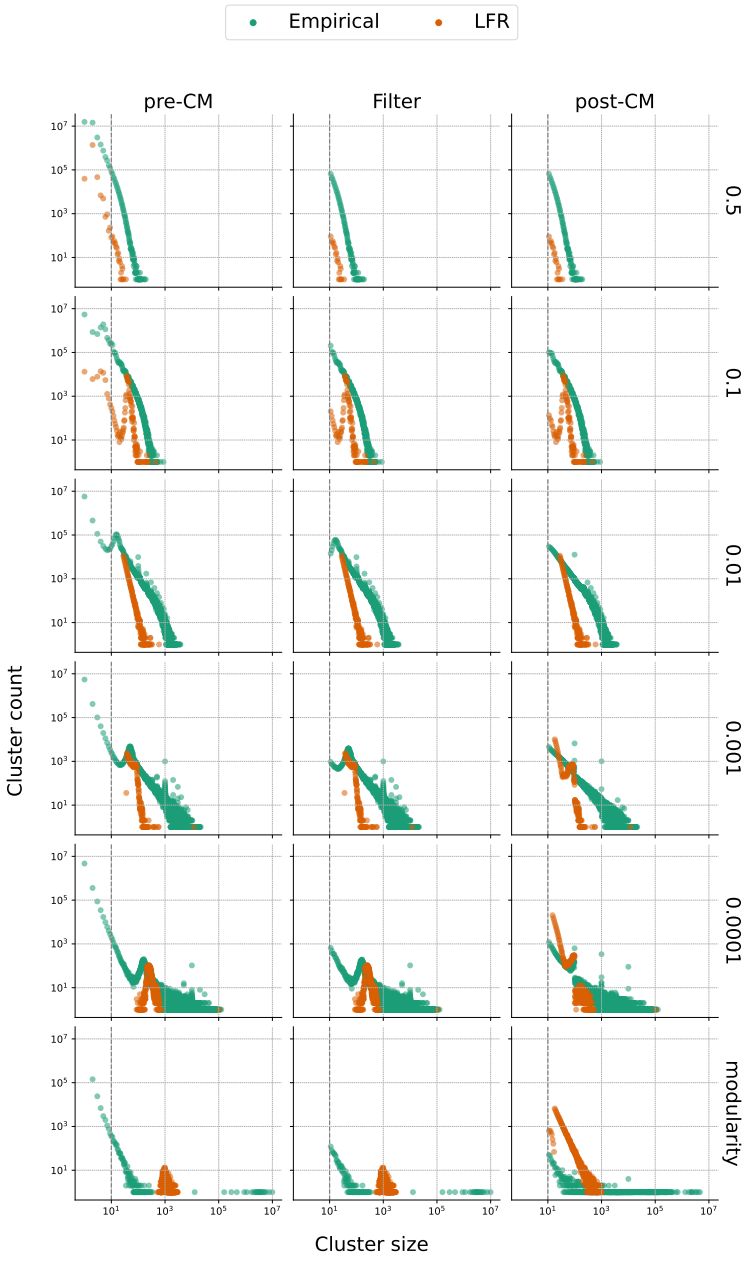

Figure D: Impact of CM-processing on cluster size distributions of Leiden clusterings of real-world and LFR networks for the Open Citations network. The real-world network has 75,025,194 nodes with an average degree of 36.35. Each row represents a different Leiden clustering method, with CPM-optimization for the top 5 rows and modularity-optimization for the bottom row. Each column represents a stage in the CM pipeline. The panels show the cluster size distributions in each step of running CM for the real-world network and its corresponding LFR graph. The axes are shown in log scale.

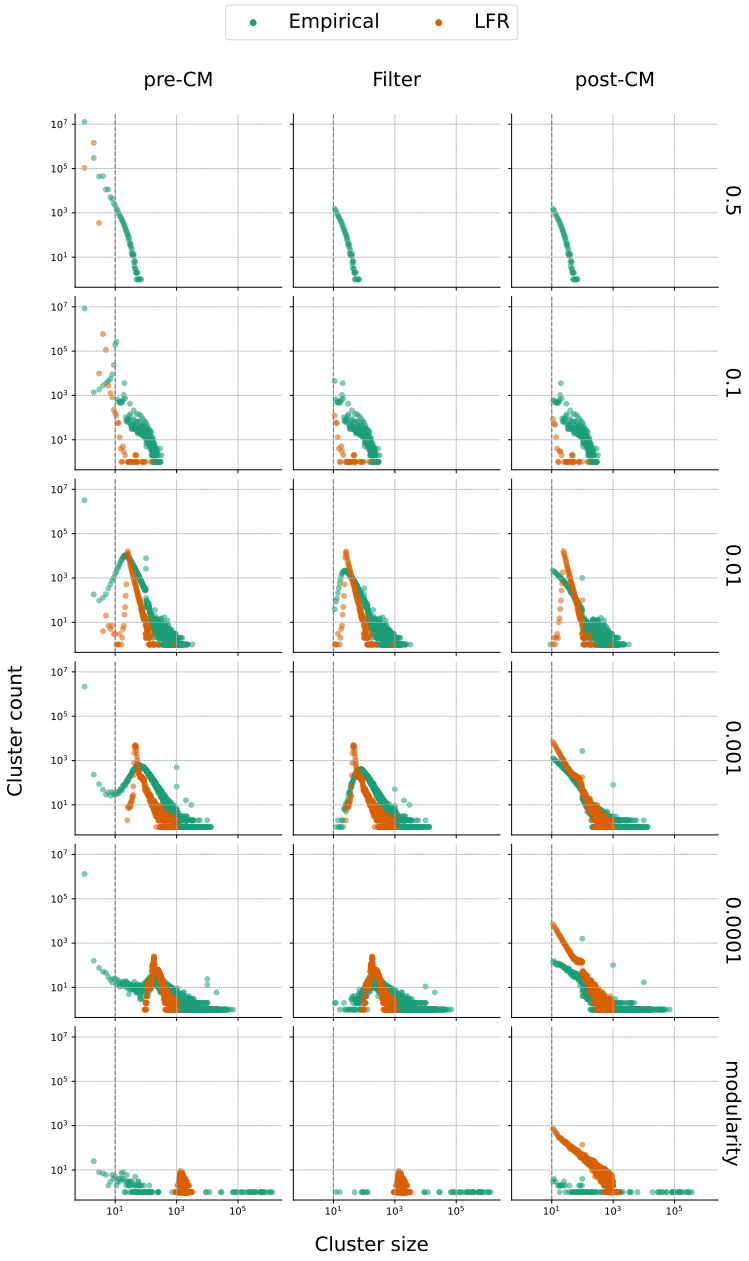

Figure E: Impact of CM-processing on cluster size distributions of Leiden clusterings of real-world and LFR networks for the CEN network. The real-world network has 13,989,436 nodes with an average degree of 13.16. Each row represents a different Leiden clustering method, with CPM-optimization for the top 5 rows and modularity-optimization for the bottom row. Each column represents a stage in the CM pipeline. The panels show the cluster size distributions in each step of running CM for the real-world network and its corresponding LFR graph. The axes are shown in log scale.

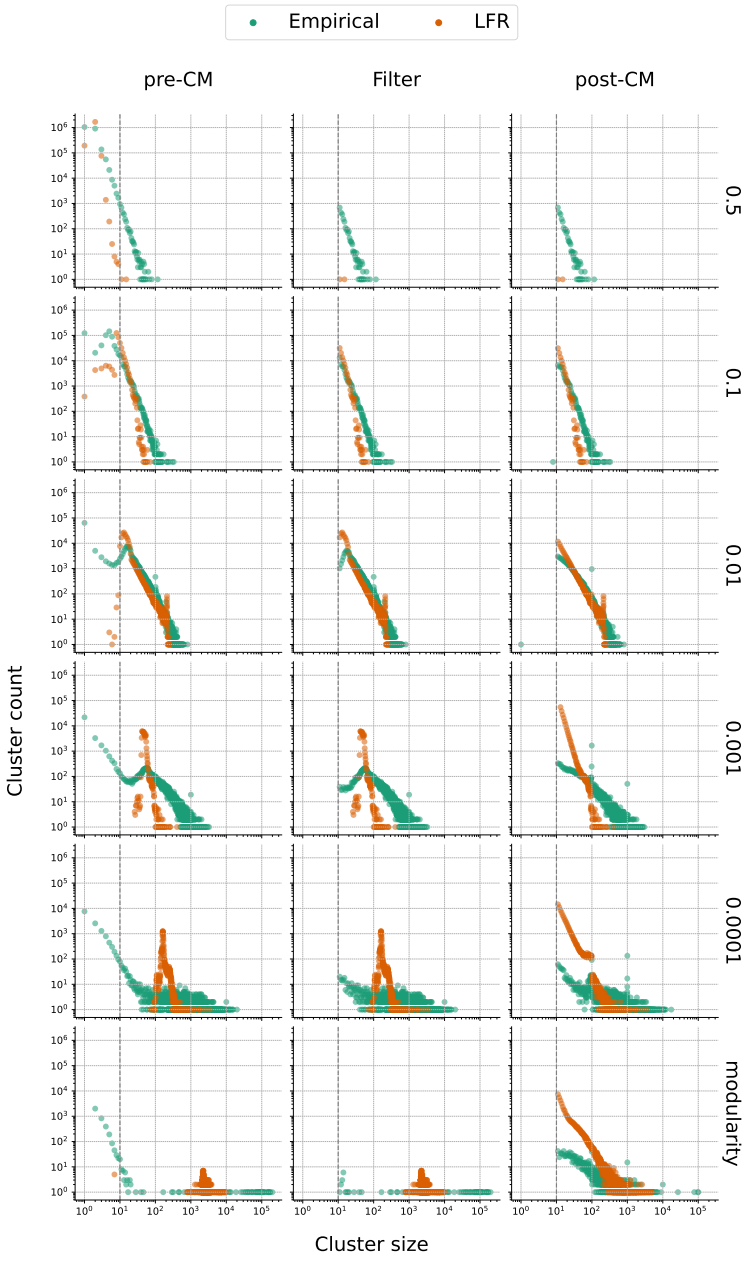

Figure F: Impact of CM-processing on cluster size distributions of Leiden clusterings of real-world and LFR networks for the cit_patents network. The real-world network has 3,774,768 nodes with an average degree of 8.75. Each row represents a different Leiden clustering method, with CPM-optimization for the top 5 rows and modularity-optimization for the bottom row. Each column represents a stage in the CM pipeline. The panels show the cluster size distributions in each step of running CM for the real-world network and its corresponding LFR graph. The axes are shown in log scale.

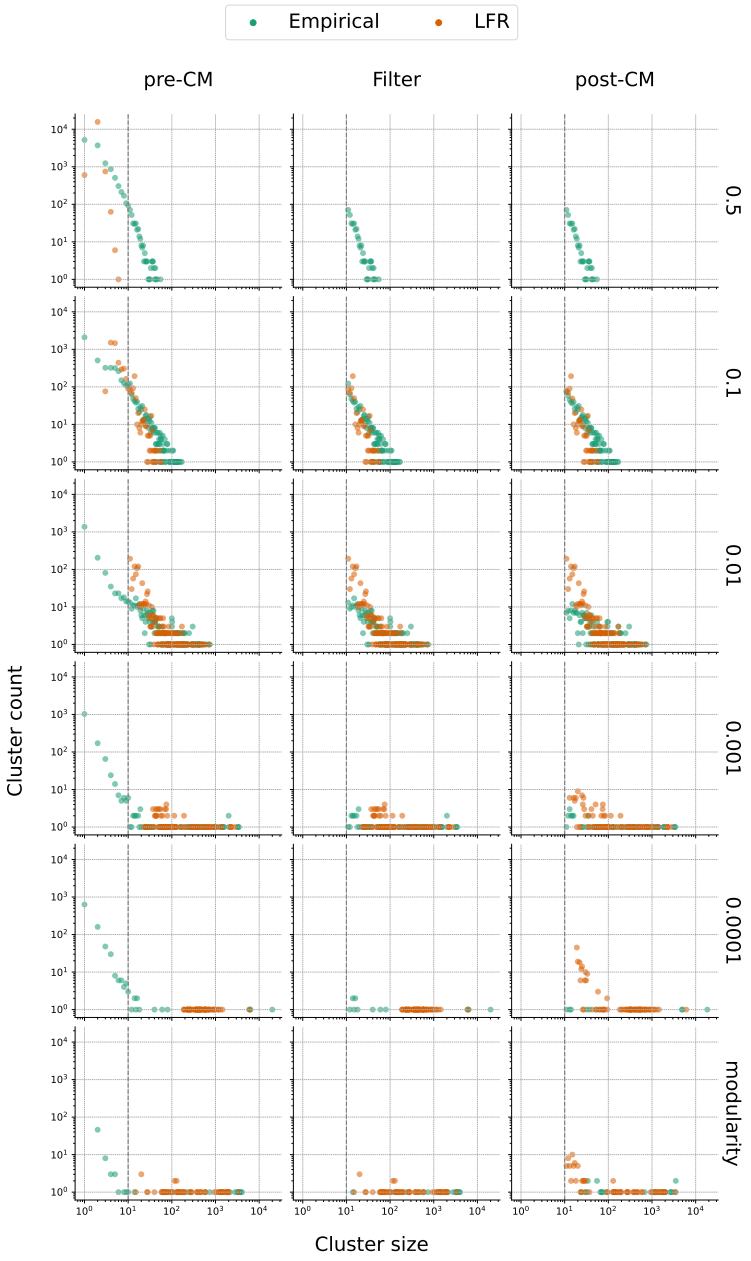

Figure G: Impact of CM-processing on cluster size distributions of Leiden clusterings of real-world and LFR networks for the cit_hepph network. The real-world network has 34,546 nodes with an average degree of 24.37. Each row represents a different Leiden clustering method, with CPM-optimization for the top 5 rows and modularity-optimization for the bottom row. Each column represents a stage in the CM pipeline. The panels show the cluster size distributions in each step of running CM for the real-world network and its corresponding LFR graph. The axes are shown in log scale.

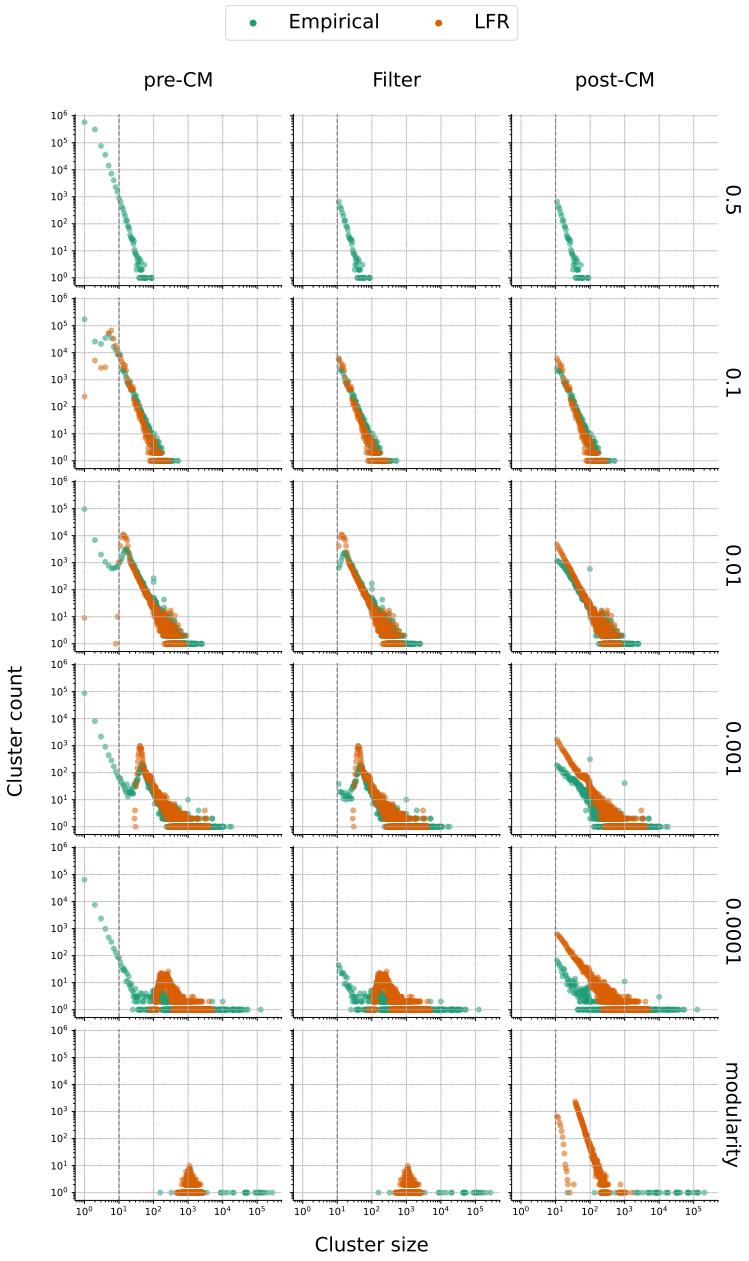

Figure H: Impact of CM-processing on cluster size distributions of Leiden clusterings of real-world and LFR networks for the wiki_topcats network. The real-world network has 1,791,489 nodes with an average degree of 28.41. Each row represents a different Leiden clustering method, with CPM-optimization for the top 5 rows and modularity-optimization for the bottom row. Each column represents a stage in the CM pipeline. The panels show the cluster size distributions in each step of running CM for the real-world network and its corresponding LFR graph. For this network, we were not able to generate a corresponding LFR graph for $r = 0.5$. The axes are shown in log scale.

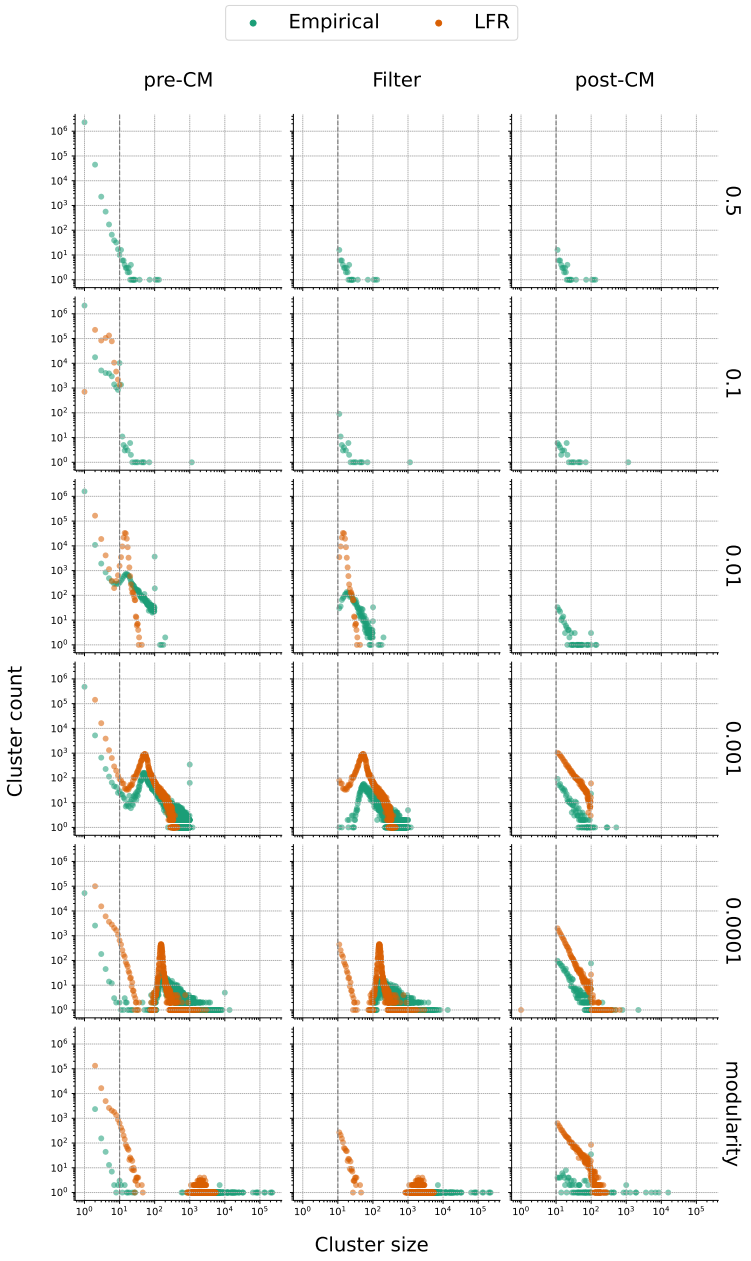

Figure I: Impact of CM-processing on cluster size distributions of Leiden clusterings of real-world and LFR networks for the wiki_talk network. The real-world network has 2,394,385 nodes with an average degree of 3.89. Each row represents a different Leiden clustering method, with CPM-optimization for the top 5 rows and modularity-optimization for the bottom row. Each column represents a stage in the CM pipeline. The panels show the cluster size distributions in each step of running CM for the real-world network and its corresponding LFR graph. For this network, we were not able to generate a corresponding LFR graph for $r = 0.5$. The axes are shown in log scale.

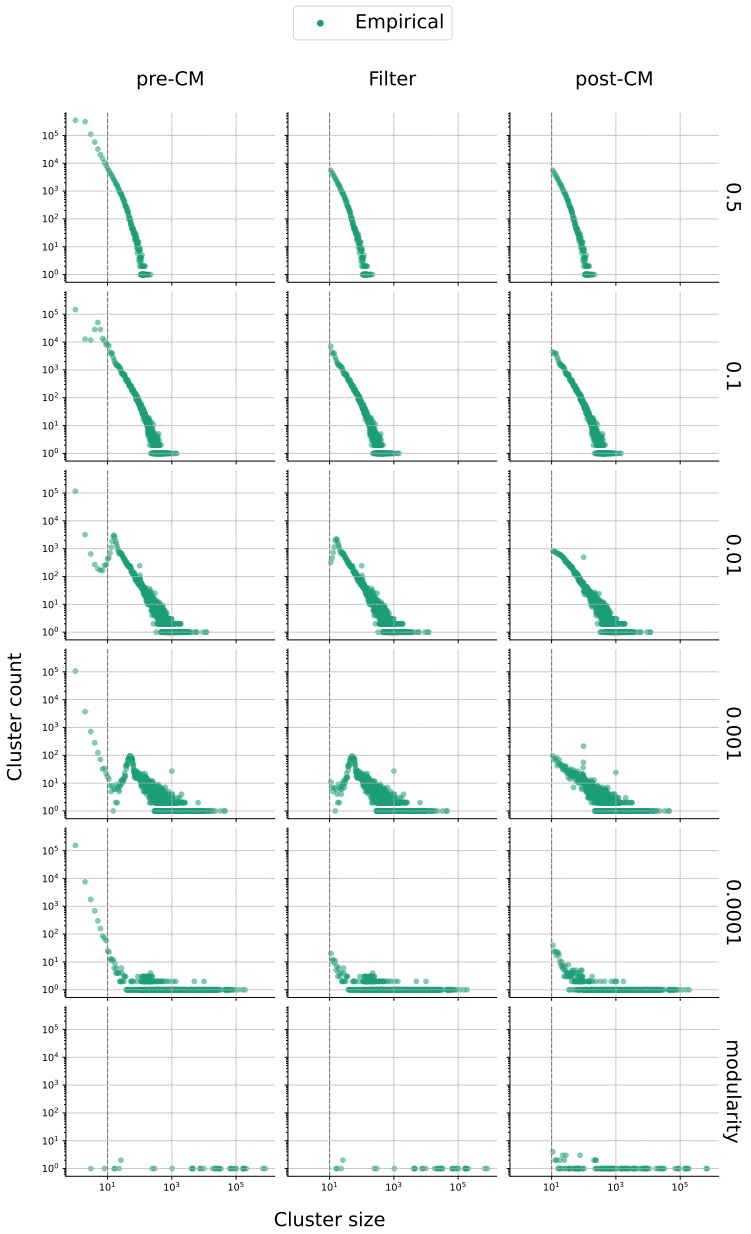

Figure J: Impact of CM-processing on cluster size distributions of Leiden clusterings of the real-world Orkut network. The real-world network has 3,072,441 nodes with an average degree of 76.28. Each row represents a different Leiden clustering method, with CPM-optimization for the top 5 rows and modularity-optimization for the bottom row. Each column represents a stage in the CM pipeline. The panels show the cluster size distributions in each step of running CM for the real-world network. For this network, we were not able to generate a corresponding LFR graph in any condition. The axes are shown in log scale.

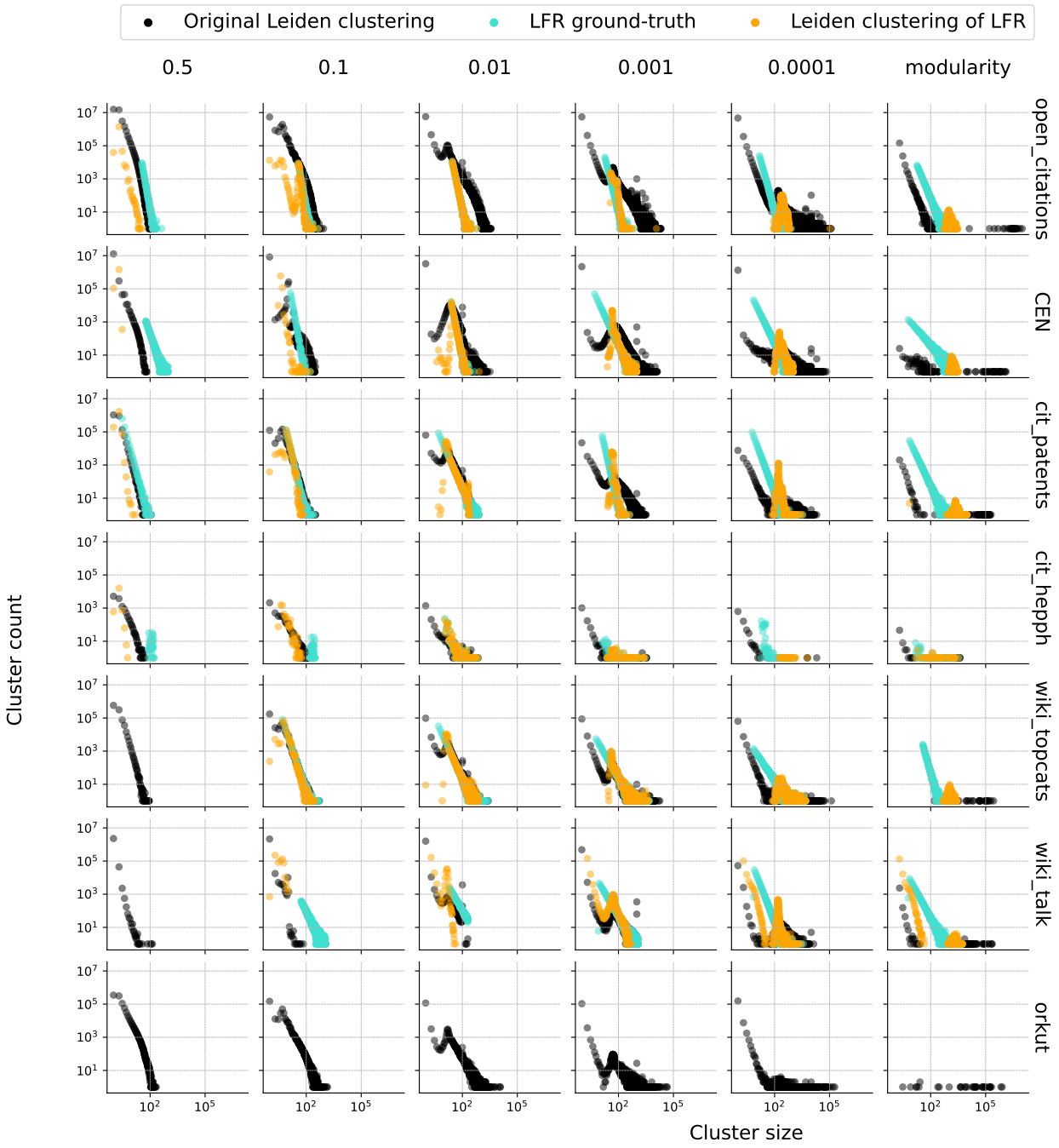

Figure K: Cluster size distribution for all LFR networks. The LFR ground-truth communities (shown in light green) are generated according to the parameters estimated from the Leiden clustering of the real-world network (shown in black), and then the LFR network is re-clustered using Leiden with the same resolution value (in orange). We were not able to generate an LFR graph for the Orkut network. The axes are shown in log scale.