MinMaxCorrelationClustering.py contains the functions used to implement our algorithm and the KMZ algorithm. These are called in all other files ending with .py

The folder Scalability Experiments refers to the three datasets considered (see Appendix G.5) to show that our algorithm scales to graphs with about 10,000 nodes. Each file ending in .py in this folder contains the experiments and analysis for one of the three datasets. The subfolder Data Files contains adjacency lists for the three datasets: LastFMAsia.csv, hepTH.csv, and hepPH.csv, needed in the .py files. This data can be found at https://snap.stanford.edu/data/feather-lastfm-social.html, https://snap.stanford.edu/data/ca-HepTh.html, and https://snap.stanford.edu/data/ca-HepPh.html, respectively.

The folder Facebook Experiments refers to the ten Facebook datasets considered. The subfolder Data Files contains adjacency lists for the ten datasets as well as the ground truth circles accompanying these. This data can be found at https://snap.stanford.edu/data/ego-Facebook.html. The file ICML_ExperimentsFacebookLarge.py is used to run experiments and analysis on the large datasets (i.e., 0, 107, 1684, 1912, 3437) and the file ICML_ExperimentsFacebookSmall.py for small datasets (i.e., 348, 414, 686, 698, 3980). Replace “FilenameGraph.csv” and “FilenameCircles.cvs” in the .py file with those of the dataset of interest, e.g., “0fb.csv” and “0fbcircles.csv”.

The folder Synthetic Experiments refers to the synthetic experiments considered. The folder ICML_PerfectWithNoise.py contains the code used for the experiments and analysis, and the file synthetic_clusters.csv contains the clusters output by the code, used to run the last part of the analysis in ICML_PerfectWithNoise.py. (If the code is run again, the content in synthetic_clusters.csv may change, since the noise is generated using randomness.)