EN.580.694: Statistical Connectomics
Final Project Report

Max Collard

May 15, 2015

How good is "perfect" clustering outside of a SBM?

Opportunity Recent work has demonstrated that there is hope for accurately grouping "similarly-connected' vertices generated by stochastic blockmodels (SBMs), as well as more general random dot product graphs (RDPGs), using mean-square error (MSE) clustering on the adjacency spectral embedding (ASE) of the graph [1].

Challenge Although asymptotically optimal for this class of generative models, it has been proposed that real-world networks, including connectomes, are generated by alternate schemes with different statistical properties than SBMs/RDPGs. The decay in the optimality of ASE-MSE clustering in this setting remains largely unexplored.

Action I applied the methods proposed in [1] to three settings:

1. A K-block SBM, where ASE-MSE clustering is asymptotically optimal. The "ground truth" here is the latent block labels assigned by the SBM scheme.

2. The Barabási-Albert (BA) preferential attachment model, starting with K cliques of equal size. Here, I decided to use "heredity" as the ground truth—that is, each of the original cliques was given a label, and each successive vertex' label was assigned by a majority vote of the labels on the vertices it connected to when added.

3. The *C. elegans* connectome, as the starting condition for a BA process of varying length. Here, the ground truth was the ASE-MSE clustering result on the *original* data: my goal was to see how the labeling of the *original* vertices is negatively affected by BA-type *perturbations* (preferential-attachment vertex additions) to the entire graph.

To simplify analysis, in all cases I let $K = 4$, and I let the dimensionality of the ASE latent space be $d = K = 4$. Performance was measured using the adjusted rand index (ARI) between the "true" and "estimated" cluster labels.

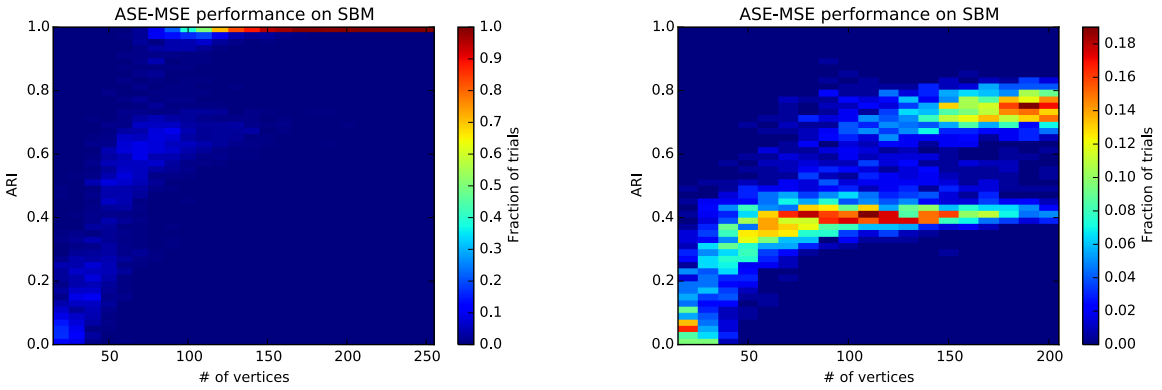

Figure 1: ASE-MSE clustering performance on data generated with a 4-block SBM, for two different parameter sets: each *column* is a histogram for a fixed number of vertices. While one converges to "perfect" clustering fairly quickly, the other does not.

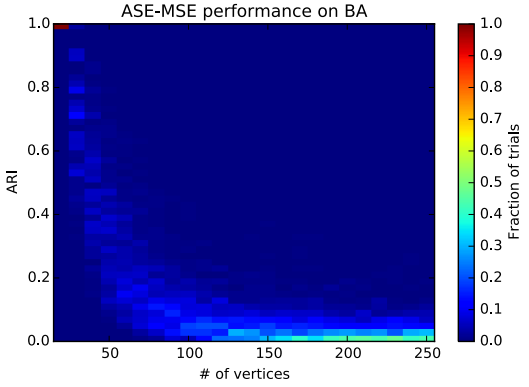

Figure 2: ASE-MSE clustering performance at identifying "heredity" on data generated using the BA model. Note the rapid decay as more vertices are added.

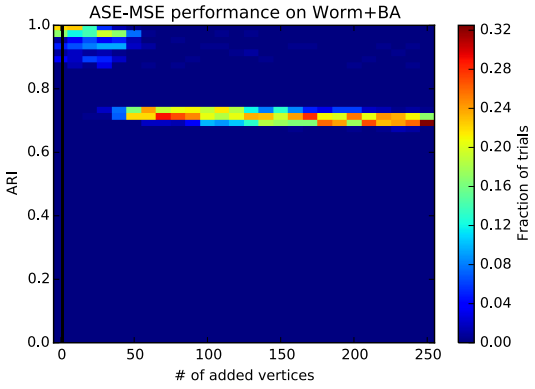

Figure 3: ASE-MSE clustering performance for the BA process started on the *C. elegans* connectome, where ground truth is the ASE-MSE labels on the unperturbed data. Performance appears to remain remarkably stable, even after adding nearly as many vertices as were in the starting dataset.

Resolution As expected, ASE-MSE clustering applied to data generated with an SBM performed well, with performance tending to increase with the number of vertices. Interestingly, it appeared as though performance tends to increase in somewhat *discrete* intervals. The specifics of these intervals, as well as the rate of convergence toward perfect accuracy, appear to depend strongly on the particular SBM parameters used. ASE-MSE clustering's ability to detect "heredity" within a BA-generated network, however, appeared to fall off very rapidly with the number of vertices added. In my view, this does not represent a shortcoming of the ASE-MSE clustering method, however; instead, it appears to show that, when the BA process "wires together" the original disconnected components, over time the information about which component "incorporated" which node is lost. Lastly, and most surprisingly, ASE-MSE clustering appears to be *remarkably* robust to perturbations of BA-type; that is, the ASE-MSE labellings appear to be affected very little by adding vertices using the BA preferential attachment scheme.

Future Work The most glaring issue with this methodology was the "heredity" vertex labelling used for the pure BA graphs: there is no reason to believe *a priori* that ASE-MSE clustering would be able to extract this information, and its failure to extract it doesn't seem intrinsically meaningful at the time of writing. Perhaps a better method would be to use the starting component that is "closest" in network distance to each vertex as that vertex' label; this would, in theory, probe a structural question about BA-generated graphs, namely whether vertices that are "close" in terms of the network topology are also "close" in terms of connectivitiy.

References

[1] Vince Lyzinski, Daniel L Sussman, Minh Tang, Avanti Athreya, Carey E Priebe, et al. Perfect clustering for stochastic blockmodel graphs via adjacency spectral embedding. *Electronic Journal of Statistics*, 8(2):2905–2922, 2014.