Stochastic Block Model

Preliminary simulations to assess different data integration techniques and their ability to estimate B

Simulation Setting: Using `graph-tool`, two graphs G_1 and G_2, representing two different sources of data on the same gene set, were generated from the same stochastic block model, with 3 true communities ($B = 3$), and connectivity matrix, $\mathbf{P}$ given as

$$\mathbf{P} = \begin{bmatrix} 0.500 & 0.025 & 0.025 \\ 0.025 & 0.500 & 0.025 \\ 0.025 & 0.025 & 0.500 \end{bmatrix}.$$

Since $\mathbf{P}$ is compound symmetric, G_1 and G_2 are drawn from an assortative planted partition model, where intra-community connectivity is dense and equal across communities, while inter-community connectivity is sparse and equal across all pairs of communities.

The graph G_1 is chosen to be (relatively) large, with $n_1 = 50$ nodes, while the graph G_2 is chosen to be small, with $n_2 = 20$ nodes. The community membership parameters $\mathbf{b}_1 = (b_{11}, \ldots, b_{1n_1})$ and $\mathbf{b}_2 = (b_{21}, \ldots, b_{2n_2})$ are chosen assigned by taking n_1 and n_2 random samples with replacement from the set $\mathbf{b} = \{0,1,2\}$, respectively. To populate G_1 and G_2 with edges, we cycle through the $(n-1)n/2$ possible pairs of vertices (v_a, v_b), and assign an edge according to a Bernoulli drawn with probability parameter P_{b_a,b_b}, where b_a and b_b are the community memberships of v_a and v_b, respectively, and P_{b_a,b_b} is the corresponding element of $\mathbf{P}$.

Three different approaches are taken to integrate G_1 and G_2 into the data-integrated graph G.

Data Integration Approach 1: A multigraph G is created by first setting $G = G_1$. Then, we add all edges of G_2 to G, where two parallel edges between the same set of nodes is allowed. We then fit the minimum description length SBM (MDL-SBM) to G, G_1, and G_2, and take note of the estimated number of blocks $\hat{B}$ estimated in graph G. We repeat this process I times.

Data Integration Approach 2: A weighted simple graph G is created by first setting $G = G_1$ and assigning G a `graph-tool` property map for edge weights, where edges that occur neither in G_1 nor G_2 are given a weight of 0, edges that occur in one of G_1 or G_2 are assigned a weight of 1, and edges that occur in both G_1 and G_2 are assigned a weight of 2. We then fit a weighted SBM (WSB) to G with a binomial model for the edge weights and record the estimated number of blocks $\hat{B}$. We repeat this process I times.

Data Integration Approach 3: A simple unweighted graph is created by setting $G = G_1$, then adding all the edged from G_2 that are not already present in G. This approach is similar to the first approach but no repeat edges are placed. We then fit a SBM to G and record the estimated number of blocks $\hat{B}$. We repeat this process I times.

Table 1: *Simulation results for each integration approach as well as for SBMs fit to G_1 and G_2 individually. The proportion of correctly specified models and the average number of clusters estimated across all iterations are shown.*

$I = 1{,}000$	$\frac{1}{I}\sum_{i=1}^{I} 1_{\hat{B}=B}$	$\frac{1}{I}\sum_{i=1}^{I} \hat{B}$
G (Approach 1)	0.70	3.38
G (Approach 2)	0.04	4.33
G (Approach 3)	***0.81***	***3.04***
G_1	0.73	3.24
G_2	0.00*	1.54

* G_2 likely does not have sufficient sample size to detect three communities.

It appears the approach 3 had the best performance in terms of correctly specifying the number of clusters.