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 ($B = 3$), under two different connectivity matrices, $\mathbf{P}_1$ and $\mathbf{P}_2$ given as

$$\mathbf{P}_1 = \begin{bmatrix} 0.500 & 0.100 & 0.100 \\ 0.100 & 0.500 & 0.100 \\ 0.100 & 0.100 & 0.500 \end{bmatrix}, \quad \mathbf{P}_2 = \begin{bmatrix} 0.300 & 0.100 & 0.100 \\ 0.100 & 0.300 & 0.100 \\ 0.100 & 0.100 & 0.300 \end{bmatrix}.$$

The graphs G_1 and G_2 have equal number of nodes $n = n_1 = n_2$, and are assigned identical community labeling parameters $\mathbf{b}_1 = (b_{11}, \ldots, b_{1n_1}) = \mathbf{b}_2 = (b_{21}, \ldots, b_{2n_2}) = \mathbf{b}$, where $\mathbf{b}$ is assigned by taking n random samples with replacement from the set $\{0,1,2\}$. 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 either $\mathbf{P}_1$ or $\mathbf{P}_2$.

Data Integration Approach: 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 can be thought of as populating G with the set union of all edges between graphs G_1 and G_2. 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 SBMs fit to G, G_1, and G_2* ***under*** *$\mathbf{P}_1$* ***and*** *$\mathbf{P}_2$, with $n = 50$. The proportion of correctly specified models and the average number of clusters estimated are shown.*

	$\mathbf{P}_1$		$\mathbf{P}_2$	
$I = 100$	$\frac{1}{I}\sum_{i=1}^{I} 1_{\hat{B}=B}$	$\frac{1}{I}\sum_{i=1}^{I} \hat{B}$	$\frac{1}{I}\sum_{i=1}^{I} 1_{\hat{B}=B}$	$\frac{1}{I}\sum_{i=1}^{I} \hat{B}$
G	***0.85***	***2.85***	***0.00***	***1.01***
G_1	0.47	2.50	0.00	1.00
G_2	0.42	2.38	0.00	1.00

Table 2: *Simulation results for SBMs fit to G, G_1, and G_2* ***under*** $\mathbf{P}_1$ ***and*** $\mathbf{P}_2$*, with $n = 100$. The proportion of correctly specified models and the average number of clusters estimated are shown.*

	$\mathbf{P}_1$		$\mathbf{P}_2$	
$I = 100$	$\frac{1}{I}\sum_{i=1}^{I} 1_{\hat{B}=B}$	$\frac{1}{I}\sum_{i=1}^{I} \hat{B}$	$\frac{1}{I}\sum_{i=1}^{I} 1_{\hat{B}=B}$	$\frac{1}{I}\sum_{i=1}^{I} \hat{B}$
G	***1.00***	***3.00***	***0.97***	***2.99***
G_1	1.00	3.00	0.23	1.70
G_2	0.99	3.01	0.19	1.76