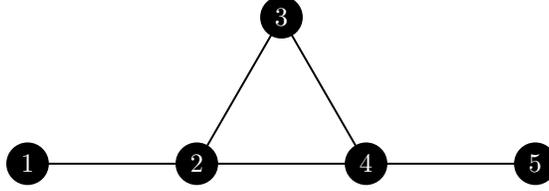


A The Glyph Algebra

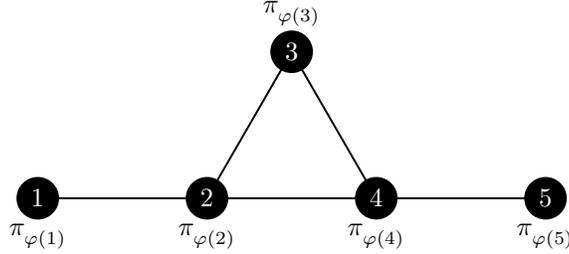
In this paper we use small glyphs as algebraic quantities. This appendix describes how to compute these quantities from the SBM, and the how they relate to the operations used in the inference method.

A.1 Subgraph Densities

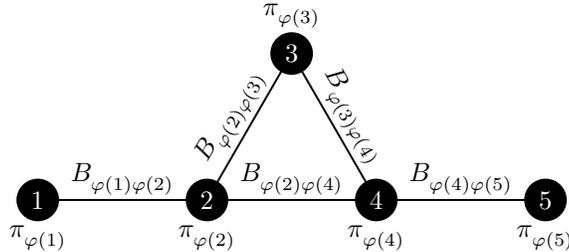
We are interested in the homomorphism density of a subgraph g in a distribution given by $\text{SBM}(\pi, \mathbf{B})$. Consider the probability that 5 randomly sampled nodes form the subgraph g shown below:



Let φ be a map from $V(g)$ to $[K]$ that assigns each node to one of the K blocks. Nodes are sampled independently from the distribution π , so the probability of such a mapping is given by the product of the corresponding entries of π :



Given one such assignment φ , the probability of seeing the subgraph g is the product of the edges probabilities, given by the corresponding entries of \mathbf{B} :



To get the density of subgraph g , we sum over all $K^{|V(g)|}$ possible maps:

$$\underbrace{\mu(g)}_{\text{homomorphism density } \mu \text{ of a subgraph } g \text{ in an SBM given by } \pi \text{ and } \mathbf{B} \text{ with } K \text{ blocks}} \Big|_{\text{SBM}(\pi, \mathbf{B})} = \underbrace{\sum_{\varphi: V(g) \rightarrow [K]}^{K^{|V(g)|}}}_{\text{sum over all maps } \varphi \text{ from vertices in } g \text{ to the } K \text{ blocks}} \left[\underbrace{\left(\prod_{i \in V(g)} \pi_{\varphi(i)} \right)}_{\text{probability of that vertex assignment}} \times \underbrace{\left(\prod_{(i,j) \in E(g)} B_{\varphi(i)\varphi(j)} \right)}_{\text{probability of the corresponding edges}} \right] \quad (35)$$

A.2 Rooted Subgraph Densities

Consider the ‘‘Feynman diagram’’ for a single edge:

$$\begin{array}{c} \textcircled{1} \xrightarrow{B_{\varphi(1)\varphi(2)}} \textcircled{2} \\ \pi_{\varphi(1)} \qquad \qquad \pi_{\varphi(2)} \end{array} \longleftrightarrow \sum_{\varphi(1) \in [K]} \sum_{\varphi(2) \in [K]} \pi_{\varphi(1)} B_{\varphi(1)\varphi(2)} \pi_{\varphi(2)}$$

Summing over all node assignments results in a scalar, namely, the edge density of the entire SBM. To represent quantities specific to individual blocks, we consider subgraphs with one of its vertices designated as the ‘‘root’’, which we represent as an open circle.

To compute these quantities, omit the \sum and π associated with the rooted vertex:

$$\begin{array}{c} \textcircled{1} \xrightarrow{B_{\varphi(1)\varphi(2)}} \textcircled{2} \\ \pi_{\varphi(2)} \end{array} \longleftrightarrow \sum_{\varphi(2) \in [K]} B_{\varphi(1)\varphi(2)} \pi_{\varphi(2)}$$

As we are no longer summing over $\varphi(1) \in [K]$, this computation results in a scalar for each of the K possible assignments of $\varphi(1)$. In other words, the result is a vector of length K (in this case, the vector \mathbf{d} of normalized degrees), indexed by the K blocks.

In general, the length- K vector associated to a rooted subgraph is given by:

$$\underbrace{\mu_k(g; v)}_{\text{homomorphism density } \mu \text{ of a subgraph } g \text{ with rooted vertex } v \text{ in block } k} \Big|_{\text{SBM}(\boldsymbol{\pi}, \mathbf{B})} = \underbrace{\sum_{\substack{\varphi: V(g) \rightarrow [K] \\ \text{s.t. } \varphi(v) = k}}}_{\text{sum over all maps } \varphi \text{ that send vertex } v \text{ to block } k} \left[\underbrace{\left(\prod_{i \in V(g) \setminus v} \pi_{\varphi(i)} \right)}_{\text{probability of that assignment of remaining vertices}} \times \underbrace{\left(\prod_{(i,j) \in E(g)} B_{\varphi(i)\varphi(j)} \right)}_{\text{probability of the corresponding edges}} \right] \quad (36)$$

A.3 Pointwise Product \equiv Gluing Graphs

Instead of drawing out the large diagrams with labels, we represent the quantities associated with rooted subgraph densities as small glyphs. For example, the vector of normalized degrees is:

$$\textcircled{1} \longleftrightarrow \mathbf{d}$$

The graph pencil method uses an operation denoted by \circ , which indicates entrywise multiplication (or Hadamard product). In the context of rooted subgraph densities, this is known as the gluing operation [21], as it corresponds to ‘‘gluing’’ together rooted subgraphs by their roots. For example, we can take powers of the vector of degree densities:

$$\begin{array}{l} \textcircled{1} \circ \textcircled{1} = \textcircled{1} \circ \textcircled{1} \longleftrightarrow \mathbf{d} \circ \mathbf{d} = \mathbf{d}^2 \\ \textcircled{1} \circ \textcircled{1} \circ \textcircled{1} = \textcircled{1} \circ \textcircled{1} \circ \textcircled{1} \longleftrightarrow \mathbf{d} \circ \mathbf{d}^2 = \mathbf{d}^3 \end{array}$$

The fact that these expressions hold for the glyphs can be seen by factoring the sum over products in equation (36). For example:

$$\begin{aligned} \textcircled{1} \circ \textcircled{1} &= \sum_{\varphi(2) \in [K]} \sum_{\varphi(3) \in [K]} B_{\varphi(1)\varphi(2)} B_{\varphi(1)\varphi(3)} \pi_{\varphi(2)} \pi_{\varphi(3)} \\ &= \left(\sum_{\varphi(2) \in [K]} B_{\varphi(1)\varphi(2)} \pi_{\varphi(2)} \right) \left(\sum_{\varphi(3) \in [K]} B_{\varphi(1)\varphi(3)} \pi_{\varphi(3)} \right) \\ &= \textcircled{1} \circ \textcircled{1} \end{aligned}$$

A.4 Inner Product \equiv ‘‘Unrooting’’ Vertices

In section A.2, we saw that when a node is the root, we omit its corresponding \sum and π in the expression. In the same way, we can interpret the filling in (or ‘‘unrooting’’) of a vertex as performing

a weighted sum:

$$\begin{array}{c} \bullet \\ \pi_{\varphi(1)} \end{array} \longleftrightarrow \sum_{\varphi(1) \in [K]} \pi_{\varphi(1)}$$

Thus, given a vector corresponding to a rooted diagram, we can take the inner product, denoted by \bullet , with π . This corresponds to “unrooting” the rooted vertex, and the result is the corresponding subgraph density for the SBM:

$$\bullet \bullet \Psi = \Psi \longleftrightarrow \pi \bullet \mathbf{d}^3 = \sum_{k=1}^K \pi_k d_k^3 = \mu(\Psi)$$

A.5 Birooted Subgraph Densities

The procedure for birooted subgraphs is much the same as for singly-rooted subgraphs, except that now there are two designated vertices for which we omit the \sum and π . This results in a K -by- K matrix, indexed by ordered pairs of the K blocks. The simplest example is the birooted edge:

$$\begin{array}{c} \textcircled{1} \\ B_{\varphi(1)\varphi(2)} \end{array} \text{---} \begin{array}{c} \textcircled{2} \\ \end{array} \longleftrightarrow B_{\varphi(1)\varphi(2)}$$

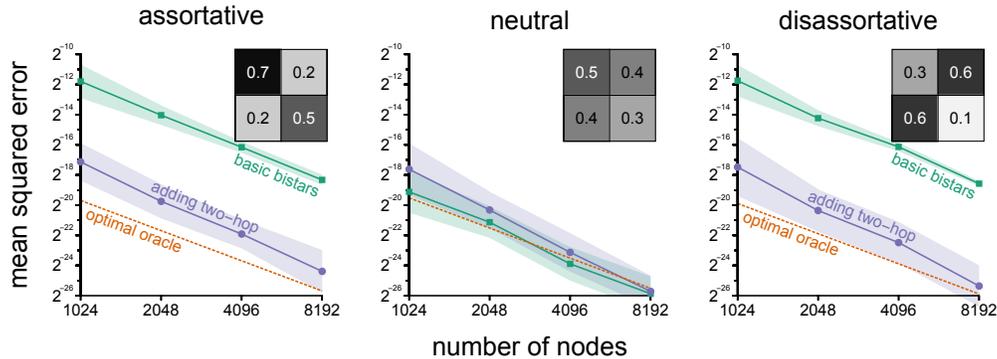
In this case, there is no sum over vertex assignments, and the only term in the product is $B_{\varphi(1)\varphi(2)}$, so this is simply the K -by- K matrix of connection probabilities:

$$\text{---} \longleftrightarrow \mathbf{B}$$

The gluing operation for birooted subgraphs is analogous to that for singly rooted subgraphs, corresponding to entrywise multiplication of K -by- K matrices. In glyph notation, each of the rooted vertices in one subgraph is glued to the corresponding rooted vertex in the other. Unrooting the two vertices requires two dot products with π .

B Simulation Details

Our graph pencil method can naturally be extended to use the “two-hop” subgraphs (section 4.3). Figure 1 (reproduced below) shows how doing so can improve the accuracy of the method in some cases (namely, those with notable (dis)assortativity):



The three figures show the results for three 2-by-2 SBMs with moderate degree separation and varying (dis)assortativity. As all three SBMs have equal-sized blocks $\pi = (0.5, 0.5)$, the gray-scaled 2-by-2 matrices displaying the connection probabilities \mathbf{B} are suggestively accurate representations of typical adjacency matrices.

From these SBMs, we sample graphs of different sizes (section 2), compute the necessary subgraph densities (section 5), and infer the parameters π and \mathbf{B} (section 4). The **green squares** represent the

basic method using bistars (section 4.2), and the purple circles represent the method that adds the two-hop subgraphs (section 4.3).

The y-axis measures the mean squared error of the inferred entries of \mathbf{B} :

$$\text{squared error} = \boldsymbol{\pi}_{\text{true}}^\top \left(\mathbf{B}_{\text{infer}} - \mathbf{B}_{\text{true}} \right)^2 \boldsymbol{\pi}_{\text{true}} \quad (37)$$

For each SBM, we simulated 16 graphs with 8192 nodes, 32 graphs with 4096 nodes, 64 graphs with 2048 nodes, and 128 graphs with 1024 nodes. Curves show the average squared error for these runs, and shading denotes ± 1 standard deviation, linearized so as to be symmetric on a log plot:

$$\left(\text{mean, stdev} \right) \longrightarrow \log(\text{mean}) \pm \frac{\text{stdev}}{\text{mean}} \quad (38)$$

The dashed orange line represents the expected squared error if one were to know the latent block assignments of the nodes:

$$\text{Var}(B_{ij}) = \begin{cases} \frac{B_{ij}(1-B_{ij})}{(n\pi_i)(n\pi_j)} & i \neq j \\ \frac{B_{ii}(1-B_{ii})}{(n\pi_i)^2/2} & i = j \end{cases} \quad (39)$$

And the dashed line shows $\boldsymbol{\pi}_{\text{true}}^\top \text{Var}(\mathbf{B}) \boldsymbol{\pi}_{\text{true}}$.