The simulated pairwise interactions of transcription factors (TFs) obtained in Chapter __ naturally induces a network, thus we can gain new insights to higher-order TF interactions and transcription pathways by clustering such network. In choosing the clustering algorithm, we need to take into consideration the multi-membership feature of the network–that is, one TF may participate in multiple complexes. This feature limits the choice of algorithms, as most common clustering algorithms assign each nodes to a single cluster. In this study we will implement an framework proposed by Ball, Newman, and Karrer [cite].

Multimembership Clustering Framework

The idea presented in [cite] is to build a generative probabilistic model for the links between nodes. Suppose there are n nodes and K clusters in the network. It's worth mentioning that in this framework, each node is associated with K parameters, denoted $\theta_{i1}, ..., \theta_{iK}$, in which θ_{iz} is interpreted as TF i's affinity to bind in cluster z. The edge between TF i and j in the network are formed according to Poisson distribution with rate $\sum_z \theta_{iz}\theta_{jz}$, and all edges are formed independently. Thus we can write the pseudo-likelihood of the network to be

$$\mathcal{L}(A, \theta) = \prod_{i,j} \left(\sum_z \theta_{iz}\theta_{jz} \right)^{A_{ij}} e^{-\sum_z \theta_{iz}\theta_{jz}}$$

and the log likelihood is

$$l(A, \theta) = \sum_{i,j} A_{ij} \log \left(\sum_z \theta_{iz}\theta_{jz} \right) - \sum_{i,j,z} \theta_{iz}\theta_{jz}$$

If we were to maximize this likelihood with respect to θ_{iz}, we will get a difficult system of non-linear equations. One way to side-step this obstacle is to introduce an auxillary variable $q_{ij}(z)$ with the constraint that $\sum_{i,j} q_{ij}(z) = 1$ and apply Jensen's inequality to get

$$\begin{aligned} l(A, \theta) &\geq \sum_{i,j} A_{ij} \sum_z q_{ij}(z) \log \left(\frac{\theta_{iz}\theta_{jz}}{q_{ij}(z)} \right) - \sum_{i,j,z} \theta_{iz}\theta_{jz} \qquad (1) \\ &= \sum_{i,j,z} \left[A_{ij} q_{ij}(z) \log \left(\frac{\theta_{iz}\theta_{jz}}{q_{ij}(z)} \right) - \theta_{iz}\theta_{jz} \right] \end{aligned}$$

This is a lower bound of the original likelihood, we will make a leap of faith and assume it is sufficient just to maximize the lower bound.

It is not hard to see that equality holdes in (1) only if

$$q_{ij}(z) = \frac{\theta_{iz}\theta_{jz}}{\sum_z \theta_{iz}\theta_{jz}} \qquad (2)$$

Thus holding θ_{iz}'s constant, the likelihood is maximized at (2).

On the other hand, if we hold $q_{ij}(z)$ constant, then θ_{iz} are maximized at

$$\begin{aligned} \theta_{iz} &= \frac{\sum_j A_{ij} q_{ij}(z)}{\sum_i \theta_{iz}} \\ \sum_i \theta_{iz} &= \frac{\sum_{i,j} A_{ij} q_{ij}(z)}{\sum_i \theta_{iz}} \\ \sum_i \theta_{iz} &= \sqrt{\sum_{i,j} A_{ij} q_{ij}(z)} \end{aligned}$$

Thus

$$\theta_{iz} = \frac{\sum_j A_{ij} q_{ij}(z)}{\sqrt{\sum_{i,j} A_{ij} q_{ij}(z)}} \qquad (3)$$

Therefore, we have a simple EM-like recursion for the following form:

1. randomly initialize $\theta_{iz}, \forall i, z$
2. update $q_{ij}\left(z\right)$ using (2)
3. update θ_{iz} using (3)
4. repeat steps 2 and 3 until convergence of the likelihood.

In the end we are interested in the final θ_{iz}'s. The result will be of the form

$$\theta = \begin{bmatrix} \theta_{11} & \theta_{12} & \cdots & \theta_{1K} \\ \theta_{21} & \theta_{22} & \cdots & \theta_{2K} \\ \theta_{31} & \theta_{32} & \cdots & \theta_{3K} \\ \vdots & & & \vdots \\ \theta_{n1} & \theta_{n2} & \cdots & \theta_{nK} \end{bmatrix}$$

We can normalize each row by the sum of each row, i.e let $\theta_i = \sum_z \theta_{iz}$, let

$$M = \begin{bmatrix} \frac{1}{\theta_1} & & & \\ & \frac{1}{\theta_2} & & \\ & & \ddots & \\ & & & \frac{1}{\theta_n} \end{bmatrix}$$

and

$$\tilde{\theta} = M\theta$$

Then entries $\tilde{\theta}_{iz}$ can be interpretated as the probability that the TF i belongs to the cluster z.