Exercises for Statistical analysis of network data – Sheet xxxxx

1. X is the data matrix. We have assumed
$$\mathrm{E}X = \boldsymbol{\alpha}\boldsymbol{\beta}^T.$$
We first calculate the in and out degrees according to
$$\mathbf{d}^{(\mathrm{in})} = X\mathbf{1} \tag{1}$$
$$\mathbf{d}^{(\mathrm{out})} = X^T\mathbf{1}. \tag{2}$$
We can see that
$$\mathrm{E}\mathbf{d}^{(\mathrm{in})} = \boldsymbol{\alpha}\boldsymbol{\beta}^T\mathbf{1} = \boldsymbol{\alpha}\|\boldsymbol{\beta}\|_1 \tag{3}$$
$$\mathrm{E}\mathbf{d}^{(\mathrm{out})} = (\boldsymbol{\alpha}\boldsymbol{\beta}^T)^T\mathbf{1} = \boldsymbol{\beta}\|\boldsymbol{\alpha}\|_1. \tag{4}$$
Note that generally $\dim\{\mathrm{E}\mathbf{d}^{(\mathrm{in})}\} \neq \dim\{\mathrm{E}\mathbf{d}^{(\mathrm{out})}\}$. We then calculate
$$\|\mathrm{E}\mathbf{d}^{(\mathrm{in})}\|_1 = \|\boldsymbol{\beta}\|_1\|\boldsymbol{\alpha}\|_1 = \|\mathrm{E}\mathbf{d}^{(\mathrm{out})}\|_1. \tag{5}$$
We therefore see that
$$\mathrm{E}\mathbf{d}^{(\mathrm{in})}/\|\mathrm{E}\mathbf{d}^{(\mathrm{in})}\|_1 = \boldsymbol{\alpha}\|\boldsymbol{\beta}\|_1/(\|\boldsymbol{\beta}\|_1\|\boldsymbol{\alpha}\|_1) = \boldsymbol{\alpha}/\|\boldsymbol{\alpha}\|_1.$$
Similarly we have that
$$\mathrm{E}\mathbf{d}^{(\mathrm{out})}/\|\mathrm{E}\mathbf{d}^{(\mathrm{out})}\|_1 = \boldsymbol{\beta}\|\boldsymbol{\alpha}\|_1/(\|\boldsymbol{\beta}\|_1\|\boldsymbol{\alpha}\|_1) = \boldsymbol{\beta}/\|\boldsymbol{\beta}\|_1.$$
One has to decide on a marginalization, this is sensible, and then we may set
$$\widehat{\boldsymbol{\alpha}} = C\mathbf{d}^{(\mathrm{in})}.$$
We now need to determine the normalization of $C > 0$. We have one more constraint, namely that $\|\boldsymbol{\alpha}\|_1 = \|\boldsymbol{\beta}\|_1$ and so we have $\|\boldsymbol{\alpha}\|_1 = \sqrt{\|\mathrm{E}\mathbf{d}^{(\mathrm{in})}\|}$ and by symmetry the same for $\boldsymbol{\beta}$. Thus we set
$$\hat{\boldsymbol{\alpha}} = \mathbf{d}^{(\mathrm{in})}/\sqrt{\|\mathrm{E}\mathbf{d}^{(\mathrm{in})}\|}.$$
The random variable $\mathbf{d}^{(\mathrm{in})}$ is a sum of independent Bernoulli random variables and a Lyapanov Central Limit Theorem can be applied. We recall that
$$d_i^{(\mathrm{in})} = \sum_j X_{ij},$$
and that each element of X is independent. Therefore by the Lyapanov Central Limit Theorem the quantity becomes Gaussian. Ditto for the out degrees.

We determine the mean and variance of $\|\mathrm{E}\mathbf{d}^{(\mathrm{in})}\|$ using direct computation, so also for the norm of the out degrees. Using Chebychev's inequality we can therefore deduce that $\|\mathrm{E}\mathbf{d}^{(\mathrm{in})}\|$ converges in probability. Informally we therefore just need to calculate the mean and variance of the in and out degrees.

As the the estimator is Gaussian this is a standard χ^2 test.

2. Suppose that for some rows $i \in I \subset \{1,\ldots,n\},\ a_i = a$ and for some columns $j \in J \subset \{1,\ldots,m\},\ b_i = b$. Denote the submatrix of X_{ij} corresponding to the cluster by X_{IJ}. In the case where we are interested only in the average of the cluster expected value, but not the individual factors a and b, and moreover, the cluster is clear in the data, the natural estimator of it is the sample average: $\bar{\theta} = (|I|\,|J|)^{-1}\sum_{i\in I, j\in J} X_{ij}$. Since the elements of X_{IJ} are all independent Bernoulli variables with mean ab and variance $ab(1-ab)$, the variance of $\bar{\theta}$ is $\mathrm{Var}(\bar{\theta}) = (|I|\,|J|)^{-1}ab(1-ab)$, of course conditional on getting the cluster membership right. Comparing this to the variance of the estimators $\hat{a}$ and $\hat{b}$ described above, and considering that $\hat{\theta}$ is again the product of $\hat{a}$ and $\hat{b}$, looks like we are much better off if we can estimate well the cluster membership than if we try to estimate the full model parameter set $\{a_i, b_j\}$.