Exercises for Statistical analysis of network data – Sheet 12

1. There is a single edge missing: in the network, there are $\binom{n}{2} = 15$ possible edges, out of which 14 are observed. The missing edge is 46. We note that $1/\log(2) = 1.44$, and $1/\log(3) = 0.91$.

Edge	A_{ij}	Shortest path distance	Neighbourhood score	Jaccard coefficient	Liben-Nowell score
12	1	1	0	0	0
13	0	2	2	2/2	$0.91 + 0.91 = 1.82$
14	0	2	1	1/3	0.91
15	0	2	1	1/3	0.91
16	1	1	0	0	0
23	1	1	0	0	0
24	1	1	0	0	0
25	0	2	1	1/4	1.44
26	0	2	2	2/4	$1.44 + 0.91 = 2.35$
34	0	2	1	1/3	0.91
35	0	2	1	1/3	0.91
36	1	1	0	0	0
45	1	1	0	0	0
46	NA	2	1	1/4	1.44
56	1	1	0	0	0

2. Shortest path distance: For this small and almost completely known network with a single connected component, there is not much information in the shortest path distance: we see that all distances are 1 or 2, and since for the unknown path, it cannot be 1. Setting a threshold between -1 and -2 would trivially predict well the states of all the observed edges, but it would also necessarily predict absence for the missing edge. A threshold could produce an edge only if set below -2, but this threshold has catastrophic performance against the sanity check on observed data: all edges observed as absent would be predicted as present.

   Neighbourhood score: Setting a threshold below 1 predicts a present link, a threshold larger than 1 predicts an absent link. However, although the idea that two nodes sharing a large swap of neighbourhood are probably linked looks in principle reasonable, it fails catastrophically on this data. Links are observed as present only if they do not share neighbourhood.

   Jaccard coefficient: Similar comments as for the neighbourhood score. Although varying the threshold we can get absent or present link, none of the thresholds is able to correctly predict what is observed.

   Liben-Nowell score: Similar comments as for the neighbourhood score. Although varying the threshold we can get absent or present link, none of the thresholds is able to correctly predict what is observed.

   However, we could obtain perfect results on the observed part of the network if we predict existing link when either of the neighbourhood, Jaccard or Liben-Nowell scores are *lower* than a selected threshold.

   The first consideration should thus be coming from the science we apply network analysis to. Should we indeed expect more likely connection between nodes if they have a large shared neighbourhood? If yes, then a method and threshold selection can be imagined based on cross-validation. If not, we need to consider what other information is available on the network or on the nodes, either from the science case of the application or as further covariates on the nodes, and build a statistical model based on those.

3. We wish to model the probability of an edge as a function of the degrees of the two end nodes. There are many different ways to specify such a model; considerations driving its choice may be taken from a background knowledge of the origin of the data. One possibility is using the product of the two degrees (together with the logit link):

$$\begin{aligned} \eta_{ij} &= \beta d_i d_j + \alpha, \\ E(A_{ij}) &= g^{-1}(\eta_{ij}), \\ g(\mu) &= \log[\mu/(1-\mu)], \end{aligned}$$

but other options are also abundant. Considering the previous exercise, this type of modelling has the advantage that we do not have to guess before modelling or to decide by cross-validation what threshold to use and whether the correlation between overlapping neighbourhoods and the existence of a link is positive or negative, which may not be trivial for larger networks. Inference can be done using the GLM framework, by maximum likelihood, pseudolikelihood or Bayesian techniques. However, here too, some thought must be put into what model to use.

4. We would have a similar problem with the three neighbourhood scoring methods as in exercise 1: the single non-zero value is for edge 23, which is observed to be absent. Any threshold value would fail a cross-validation. We might think in principle to set up a logistic regression model, but the amount of data (six possible edges, of which only four is observed, and only four nodes without further information) does not make an inference reasonable.