Review report for Comparison of Cross-Validation Methods for Stochastic Block Model

Jining Qin

February 19, 2016

1 Summary

There is currently no methods for model selection in network modeling. We consider achieving that by using cross-validation. Cross-validation has been used in assessing model fits among latent variable models. There's also a variant of it called the network cross-validation. We can show that cross-validation is better than BIC. There are some model selection methods specific to stochastic block models. And there are some community detection methods we are going to consider. We will compare the performance of the methods using simulated data.

2 General comment

I think the big picture isn't stated clearly enough in this introduction. There are several paragraphs whose purpose I'm not sure about. I'm not familiar with stochastic block models, but I don't understand what the methods for SBMs are for. Are these methods based on which you will develop cross-validation model selection method, or are they methods whose performance you are going to compare using cross-validation? My guess is in the SBM setting model selection would have some specific sense which you probably mentioned, but I don't know the language so I'm not quite sure what is going on. It would be helpful to have one or two sentences like 'under the SBM setting, model selection becomes picking something something". This ambiguity is also the reason I'm not very happy with the summary in the last paragraph "we compare the performance of * all of the methods described above* ", since I'm not sure whether these are all comparable methods.

In general I think the contents are OK but there needs to be some summarizing and transitioning sentences putting the contents in the big picture and let readers know what they are currently reading fits in.

3 Specific comments

Second paragraph, second sentence: I'm not sure whether probability estimates and tie values are comparable. It might be better as 'tie value prediction based on probability estimates'. And I'm not sure whether it's common to say 'loss between A and B'.

Second paragraph, third sentence: could be better as "Cross-validation has been used before in Hoff (2008) to compare model fits among different latent variable models."

Page 2, first paragraph, second sentence: I'm not sure what purpose this sentence serve. And I'm not sure what you mean exactly. The BIC should depend on the log-likelihood and the number of variables. So do we simply need the number of variables? I cannot quite get it.