Identifying communities in networks

Stéphanie van der Pas
svdpas@math.leidenuniv.nl

Johannes Schmidt-Hieber
schmidthieberaj@math.leidenuniv.nl

There has been a lot of interest recently in identification of small subgraphs with high connectivity in a large graph. In the language of social networks, one might think of identifying social groups. Another application is to find a set of web pages discussing similar topics based on the hyperlinks on those web pages to other web pages.

The simplest model with one group can be formulated as follows. We have a set of vertices V and a subset $S \subset V$, the "community". Moreover, we have two probabilities p and q satisfying $0 < q < p < 1$. Two vertices are independently connected by an edge with probability p if both edges are in S and with probability q otherwise. The statistical problem is to estimate/reconstruct the set S and the probabilities p, q from observing the graph.

In terms of social networks, each vertex represents one person and an edge means that they are connected. The persons who are in S are more likely to be connected to each other ($p > q$) and therefore are in one group. What we can observe are all the connections but we do not know the set S and thus have to use the data in order to reconstruct it.

In this bachelor thesis project, we study estimation and confidence statements for the unknown quantities S, p, q. It is important to find procedures that allow for fast computation in order to be applicable to large datasets. The model presented above is a toy model. In a second part of the project we will generalize our methods to more general settings. Possible extensions are to study what happens if there are several subgraphs with higher connectivity, and how to find the number of communities if we do not know this number in advance. Finally, we might consider other graph structures with non-independent edges.

The student will start out by reading some papers on this topic, and compare methods proposed in the literature, applying them to the toy model. This will include implementing the procedures (in R or Matlab) and testing them on simulated and real data. When the student has become familiar with the model, we will study one of the extensions mentioned above, or another one, depending on the student's interests.

References

[1] A. Channarond, J.-J. Daudin and S. Robin (2012). Classification and estimation in the stochastic blockmodel based on the empirical degrees. *Electronic Journal of Statistics* **6**, 2574-2601.

[2] P.J. Bickel and A. Chen (2009). A nonparametric view of network models and Newman-Girvan and other modularities. *Proceedings of the National Academy of Sciences of the United States of America* **106** (50), 21068-21073.

[3] M.E.J. Newman and M. Girvan (2004). Finding and evaluating community structure in networks. *Physical Review E* **69** 026113.

[4] T.A.B. Snijders and K. Nowicki (1997). Estimation and prediction for stochastic blockmodels for graphs with laten block structure. *Journal of Classification* **14** 75-100.