Final Assignment, 36-720

Fall 2016

Due at 11:59 pm on Thursday, 20 October

This assignment is rather more free-form than the previous two.

The data file `http://www-personal.umich.edu/~mejn/netdata/cond-mat-2003.zip` contains a weighted network of co-authorships among scientists posting papers to the condensed matter section of arxiv.org, from January 1995 (when cond-mat opened) through 30 June 2003[1]. You will build and validate a model of this network, and present your findings in a scientific report.

You will need to chose an appropriate model, estimate it, analyze whether the estimated model provides a good fit to the data, and describe what the model's findings tell us about this network. Any of the models we have discussed during class may be appropriate; if you would rather use something else, feel free to do so, so long as (1) it is a model for network structure which can be fit to this data, and (2) you can explain in the report how the model works, and why you think it appropriate.

Exploratory data analysis: You can, and probably should, perform EDA on the data set. However, you do not need to include this EDA in the report. You *should* include the EDA if it bears on your choice of model, on your evaluation of the model, or on your conclusions about what the model tells us about the network.

Description of the model: Your report should include a description of the model (or models) you have chosen to work with. Assume the reader of the report is someone who made it through the first day of this class (and so knows about statistical models, and some basic jargon about graphs), but is not familiar with any statistical models of networks. Your description of the model should also explain why you think the model is appropriate for this data. Try to be as concrete as possible about what *scientific* (not statistical) question the model would let us answer. Detailed descriptions of algorithms for parameter estimation are not required.

Model assessment: You must not only estimate your chosen model, but also assess whether or not the estimated model fits the data. If the model does not fit the data perfectly, you should present evidence about the severity of the mis-fits, and discuss their importance (or lack thereof). How you assess goodness-of-fit is up to you, but you should explain to the reader why you are assessing the model in the way that you do.

[1] This is an update of the data set published in Newman (2001c), and further analyzed in Newman (2001a,b).

Data preparation: It may be appropriate, or even necessary, to do some preliminary work on the data before estimating your favored model. This may or may not including changing directed or weighted edges to undirected or unweighted ones, restricting attention to the largest connected component, sampling to use only a computationally-tractable fraction of the graph, etc. If you take such steps, your report should discuss them, and their justification. (Merely translating the data from one file format to another does not require discussion.)

Format: Your report is no more than 10 pages, including any figures and tables, but excluding references (if any). There is no prescribed series of sections[2], but the text is clearly divided into logically-organized sections. Plots are carefully labeled, with informative and legible titles, axis labels, and (if called for) sub-titles and legends; they are placed near the relevant text. All quantitative and mathematical claims are supported by appropriate derivations, included in the text, or calculations in code, or appropriate citations to the scientific literature. Numerical results are reported to appropriate precision. Code is properly integrated with a tool like R Markdown or knitr, and both the knitted file and the source file are submitted. The code is indented, commented, and uses meaningful names. All code is relevant to the text; there are no dangling or useless commands.

Submit your compiled report as a PDF, and the source files used to generate it. Upload everything to Blackboard; do not submit by e-mail.

References

Newman, Mark E. J. (2001a). "Scientific collaboration networks: I. Network construction and fundamental results." *Physical Review E*, **64**: 016131. URL `http://www-personal.umich.edu/~mejn/papers/016131.pdf`. doi:10.1103/PhysRevE.64.016131.

— (2001b). "Scientific collaboration networks: II. Shortest paths, weighted networks, and centrality." *Physical Review E*, **64**: 016132. URL `http://www-personal.umich.edu/~mejn/papers/016132.pdf`. doi:10.1103/PhysRevE.64.016132. Erratum: Newman (2006).

— (2001c). "The structure of scientific collaboration networks." *Proceedings of the National Academy of Sciences (USA)*, **98**: 404–409. URL `https://arxiv.org/abs/cond-mat/0007214`. doi:10.1073/pnas.98.2.404.

— (2006). "Erratum: Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality [Phys. Rev. E 64, 016132 (2001)]." *Physical Review E*, **73**: 039906. doi:10.1103/PhysRevE.73.039906.

[2] But something like "Introduction — Data — Model Description — Model Evaluation — Results and Conclusions" is reasonable.