Identifying Community Structures from Network Data via Maximum Likelihood Methods: Data and Algorithm Descriptions.

Jernej Čopič, Matthew O. Jackson, and Alan Kirman *

December 17, 2006

Abstract

This document contains a brief description of the supporting material for "Identifying Community Structures from Network Data via Maximum Likelihood Methods", by Čopič, Jackson, and Kirman, December 2006. The document briefly describes the data files and the Mathematica file of the implementation of the simulation algorithm. All of these files are freely available at
http://www.hss.caltech.edu/∼jernej and http://www.stanford.edu/∼jacksonm

There are four documents available:

(1) "algorithmfinal121606.nb" – a Mathematica notebook file.

(2) "all.csv" – a comma-separated excel file (mac version, so the separator is a comma, if using a pc, first open in text editor and substitute commas with semi-colons).

*Čopič is at the Cowles Foundation at Yale University, email: jernej.copic@yale.edu. Jackson is at the Department of Economics, Stanford University, Stanford, California 94305-6072, USA, email: jacksonm@stanford.edu , web site: http://www.stanford.edu/∼jacksonm/ . Kirman is at GREQAM, Université de la Mediterranée, 2 rue de la Charité, 13236 Marseille CEDEX 02, France, email: kirman@ehess.cnrs-mrs.fr . We gratefully acknowledge financial support under NSF grant SES-0316493, as well as from the Guggenheim Foundation, the Center for Advanced Studies in the Behavioral Sciences, the Lee Center for Advanced Networking and the Social Information Sciences Laboratory at Caltech. We thank Bhaskar Dutta, Matteo Marsili, and participants of the ISS seminar at Cornell for helpful discussions; and Rik Pieters and Hans Baumgartner for making their data available.

(3) “workbook3.txt" – a txt table.

(4) “summary.xls" – an Excel file.

File (1) contains the commands, along with their precise descriptions, needed to perform a data analysis as described in the paper. Be sure to read the comments carefully before using it. Some of the commands use specific simulation parameters - if you wish to change those, you need to do it manually in the commands. It is not difficult, but must be done carefully.

File (2) contains the data on cross-citations between journals that are analyzed in the paper. We thank Rik Pieters and Hans Baumgartner for making these data available to us.

File (3) contains data on maximum possible interactions between the journals as derived from the number of articles in each journal for one year. We used 2003 as the year, since for that year these data were available for all the journals on sci net. As mentioned in the paper, this is only a proxy for the cross citations for the years in the data set on cross citations “all.csv". Since cross citations are for the period of 3 years, while (3) contains maximal number of citations for each pair for the period of 1 year, the matrix in (3) has to be multiplied by 9 when used in the methods. We remark that multiplying by 9 will only affect estimation of probabilities, and the likelihood numbers, but not the optimal groupings or significance tests since this is only a scaling factor.

File (4) contains a summary of the two data sets.