Regional and Business Studies (2009) Vol 1 No 2, 39-48
Kaposvár University, Faculty of Economic Science, Kaposvár

THE MEANS OF NETWORK ANALYSIS PROVIDED BY THE R ENVIRONMENT

György KÖVÉR, Eleonóra STETTNER
Kaposvár University, Faculty of Economic Sciences, Department of Mathematics and Physics,
H-7400 Kaposvár, Guba S. u. 40.

ABSTRACT

The R open source software package is especially suitable for mathematical and statistical data analysis. It is highly flexible and by the means of packages very customizable. The theory of graphs provides practical solutions to very different problems of everyday life. Several efforts have been made to develop packages in the R environment to implement the means of the graph theory, such as the graph, igraph, networksis, netstat packages. In this paper a comparison of the aims of the different packages is undertaken. Several packages provide exact descriptions of graphs, while others use sequential importance sampling. The enrolment data of Kaposvár University from the year 2007 were processed by sequential importance sampling using the networksis package. We concluded that the study programmes offered by the faculties of Kaposvár University do not compete with each other. The analysis was done by separating the study programmes according to faculties. Further analysis might be needed to examine all the programmes pooled together.
Keywords: Network, R, sequential sampling

INTRODUCTION

The open source R environment (freely available under the GNU General Public License) supervised by the R Development Core Team (http://www.R-project.org) is a software environment for statistical computing and graphics. The development of R was initiated by *Ihaka and Gentleman* (1996).

As R lacks a general purpose graphical user interface, most of its functionality can be activated only by the command line interface. A Hungarian introduction to the command line interface is written by *Norbert Solymosi* (2005).

The results of the graph theory have been applied by several other sciences like social sciences, economics, engineering, etc. building complex theoretical or experimental models. One of the main applications for this is network analysis.

The R environment provides several program packages based on the graph theory to the researchers, which are suitable for network analysis. On the official R web page the (http://cran.at.r-project.org/) individual packages are grouped by functionality. Among others in the network analysis group the diagram, dynamicGraph, giGraph, graph, gRbase, igraph, matgraph, network, RBGL and Rgraphviz packages can be found. Some of them are general purpose packages others are more specialized. Almost all of them provide classes for data storage suiting to the purpose of the package. The graphs handled by those packages can contain millions of edges, vertices. The packages usually support some kind of visualization, some of them are even interactive.

The package *igraph* authored by *Gábor Csárdi* (2009) is one of the general purpose graph packages. The data input is expected to be done by the R environment and stored in a data frame object. The pre-processed data transformed into an object of *graph* class are suitable to be used in the area of the graph theory. The package *igraph* can also be used to generate the famous graph examples of the literature.

The package suit *statnet* (*Handcock et al.*, 2008) integrates the functionality of several other packages and includes classes to store data representing graphs allowing them to be visualised, thus they are suitable for the analysis of a variety of network data including dynamic networks.

The package *networksis* (*Admiraal and Handcock,* 2008) based on the statnet package provides the means to simulate bipartite graphs using the method of the sequential importance sampling.

In this paper the authors apply the package *networksis,* part of the statnet suit, to analyse bipartite graphs representing enrolment data of Kaposvár University to determine if the study programmes of the faculties are competitors of each other attracting applicants from secondary schools.

MATERIAL AND METHODS

After pre-processing the enrolment data of Kaposvár University in 2007 four data tables were constructed containing the relationship between the study programmes offered by the four faculties of the university and the secondary schools of the first-year students. *Table 1* shows only a part of the whole data table of the relationship of the five study programmes of the Faculty of Animal Sciences and the 108 secondary schools. 0 denotes that there is no student enrolled from a certain school to a certain study programme, 1 denotes otherwise. A summary of study programmes and secondary schools can be found in *Table 2*.

The data processing has been done by the networksis package of the R environment. The networksis package is designed to produce random graphs using sequential importance sampling. The degrees of the vertices of the sampled graphs required to be equal to the degrees of the vertices of the observed graph. In our case it means that we produce graphs representing different enrolment structures but no difference in the number of secondary schools belonging to a specific study programme is allowed.

Table 1

Relation of study programmes and secondary schools at the Faculty of Animal Sciences

	School 1	School 2	…	School 108
Study programme 1	1	0		1
Study programme 2	1	1		0
Study programme 3	1	0		0
Study programme 4	0	0		1
Study programme 5	0	1		0

Table 2

Summary of programmes and secondary schools at the four faculties

Faculty	Study programmes	Schools
Faculty of Animal Sciences	5	108
Faculty of Economic Sciences	2	51
Faculty of Arts	8	70
Faculty of Pedagogy	13	173

Plotting the observed data as graphs the network package of the R environment was used.

RESULTS AND DISCUSSION

The relationship between a university faculty and the secondary schools can be represented by a bipartite graph. The Biparted graphs illustrate the relationship between the four university faculties and the secondary schools can be seen in *Figure 1, Figure 2, Figure 3* and *Figure 4*. The very simple structure of the graph of the Faculty of Economic Sciences is explained by the low number of study programmes. The complexity of the graphs increases according to the number of study programmes offered.

Figure 1

Relation between study programmes and secondary schools at the Faculty of Animal Sciences

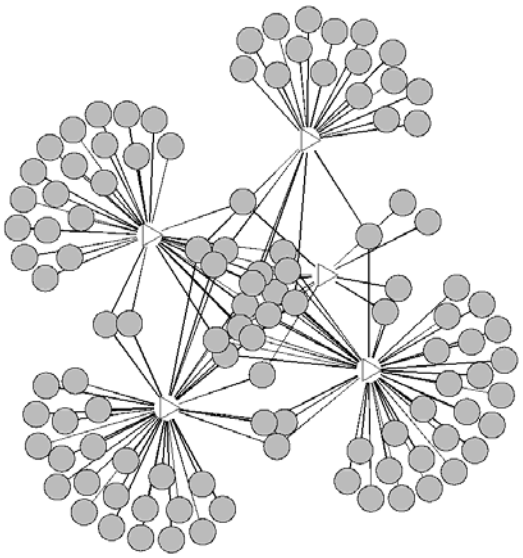


Figure 2

Relation between study programmes and secondary schools at the Faculty of Economic Sciences

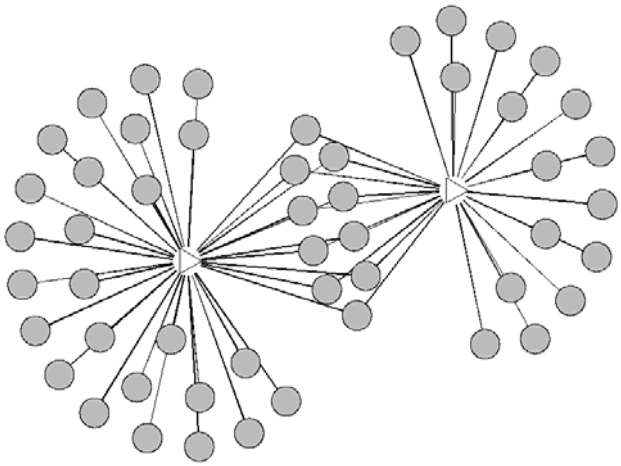


Figure 3

Relation between study programmes and secondary schools at the Faculty of Arts

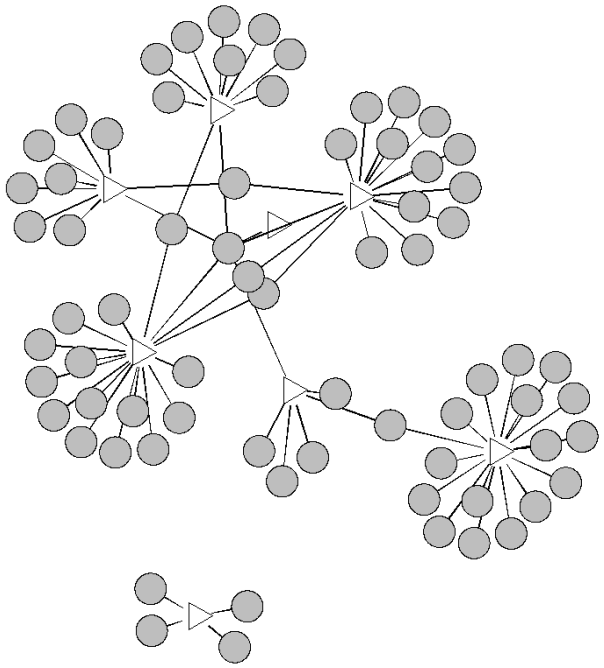
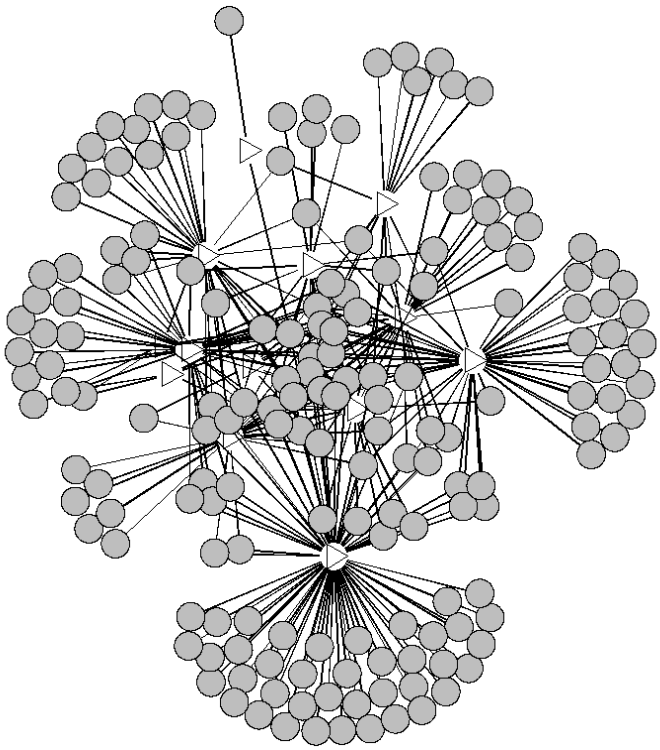


Figure 4

Relation between study programmes and secondary schools at the Faculty of Pedagogy



Estimating the size of the graph space

The graph space is defined as a set of graphs in interest. The size of a graph space is the exact enumeration of all graphs of the graph space. To calculate the total number of unique graphs meeting a certain condition can be time consuming even if the size of the gaphs is moderate.

Here we focus on enumerating the graph space where the degrees of the vertices are the same as the observed graph of enrolment of the four faculties. The graph space can be randomly sampled using the *simulate* command of the package *networksis*. To execute the *simulate* command the required sample size and the matrix representing the observed graph should be given. The estimated graph space size and its standard error can be calculated after the sampling process. Using a sequential importance sample of size 100,000 the size of the four graph spaces was estimated (*Table 3*).

The randomly generated graphs have different importance which can be expressed as the inverse of the probability of the generated graph. Using a smaller sample of size 1000 the histogram of the inverse probabilities is plotted in *Figure 5* and *Figure 6*. The sample weights approximate lognormal distribution.

Table 3

The estimated size of the graph space and standard error at the four faculties

Faculty	Size	Standard error
Faculty of Animal Sciences	1.645194e+69	5.628353e+66
Faculty of Economic Sciences	76829582426	239310495
Faculty of Arts	7.223222e+53	1.659336e+51
Faculty of Pedagogy	7.817196e+199	4.157507e+197

Figure 5

Importance weights
at Faculty of Animal Sciences and Faculty of Economic Sciences

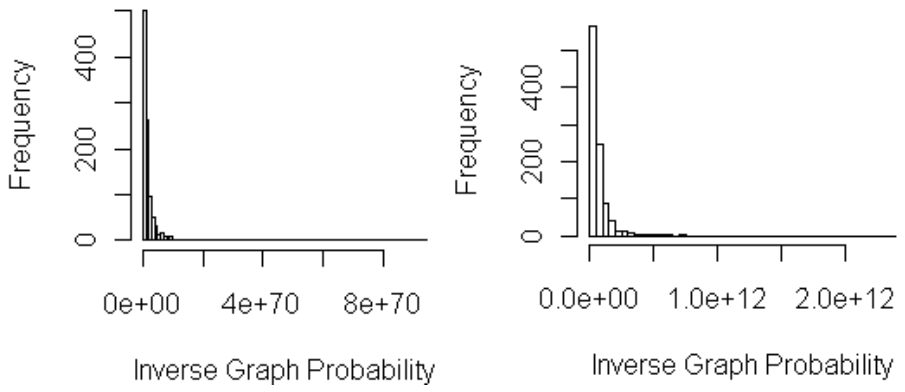
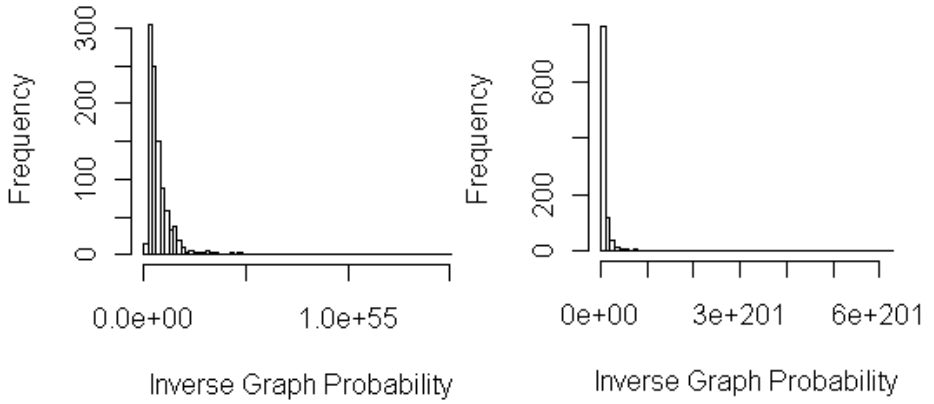


Figure 6

Importance weights at Faculty of Arts and Faculty of Pedagogy



Competition among programme studies

Roberts and Stone (1990) proposed a test statistic to test that there is a competition for food sources among certain finch species coexisting on an island. The test statistic applied to our present problem is given by $\overline{S}^2$, where *m* is the number of the programme studies of a faculty, s_{ij} is the number of those programme studies which related to secondary schools *i* and *j*. The value of $\overline{S}^2$ can be considered as the intensity of the competition.

$$\overline{S}^2 = \frac{1}{m(m-1)} \sum_{i \neq j} s_{ij}^2 \qquad (1)$$

The unknown null distribution of the test statistic $\overline{S}^2$ can be approximated by sampling the graph space. The probability of a randomly sampled graph is also calculated by the package networksis. We generated 1000 random graphs to approximate the distribution of $\overline{S}^2$. The distribution of $\overline{S}^2$ can be seen in *Figure 7*, *Figure 8*, *Figure 9* and *Figure 10* corresponding to the four faculties. Considering the Faculty of Economic Sciences the calculated values of $\overline{S}^2$ are meaningless due to the low number of study programmes. The values of $\overline{S}^2$ calculated on the base of the observed data are given in *Table 4* and represented by the thin vertical line in *Figure 7, Figure 8, Figure 9* and *Figure 10*. We also determined the number of the generated graphs with a value of $\overline{S}^2$ higher than the observed one (*Table 5*). After considering *Table 4* and *Table 5* and *Figures 7, Figure 9* and *Figure 10* we have no reason to believe that the values of $\overline{S}^2$ calculated on the base of the observed data are unexpectedly high. Our results indicate that there is no competition among study programmes after examining the faculties independently.

Figure 7

Distribution of the test statistic $\overline{S}^2$, Faculty of Animal Sciences

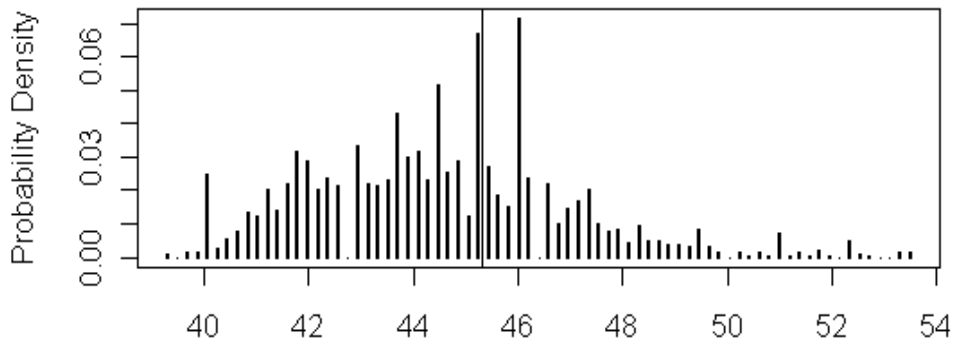


Figure 8

Distribution of the test statistic $\overline{S}^2$, Faculty of Economic Sciences

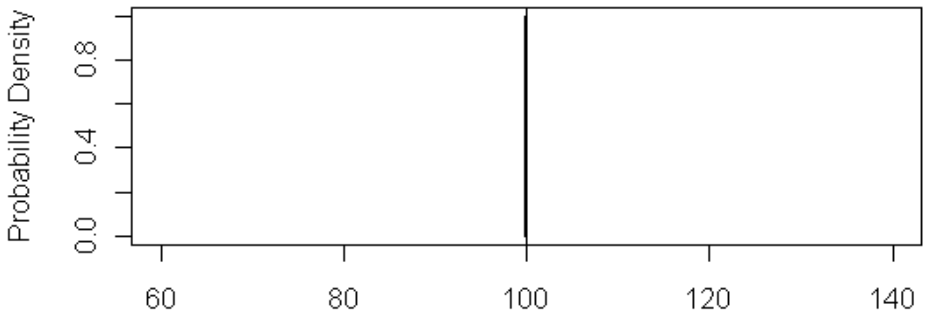


Figure 9

Distribution of the test statistic $\overline{S}^2$, Faculty of Arts

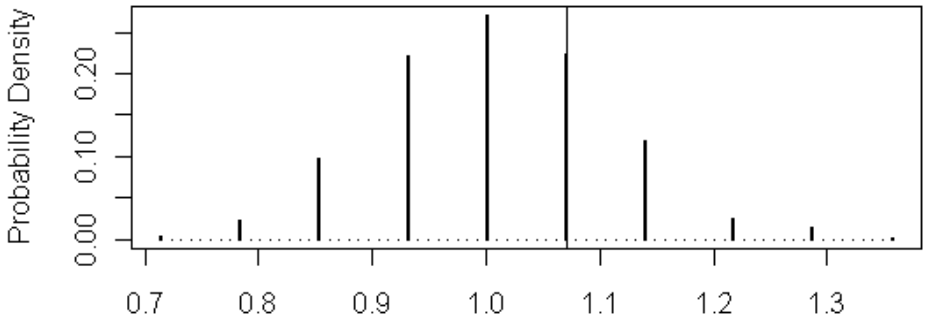


Figure 10

Distribution of the test statistic $\overline{S}^2$, Faculty of Pedagogy

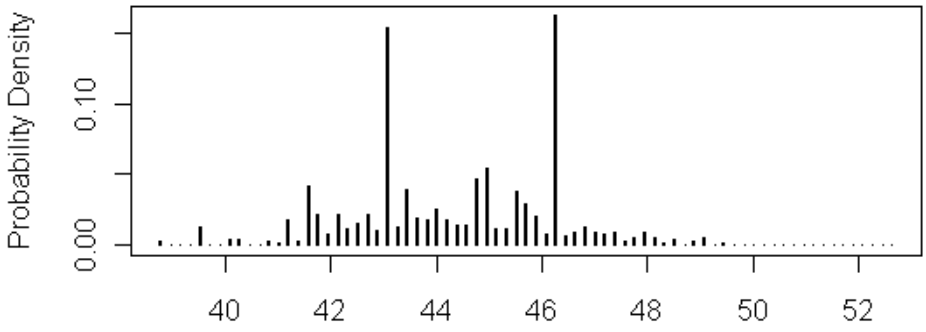


Table 4

Value of $\overline{S}^2$ statistic at the four faculties

Faculty	$\overline{S}^2$
Faculty of Animal Sciences	45.30
Faculty of Economic Sciences	100.00
Faculty of Arts	1.071429
Faculty of Pedagogy	42.69231

Table 5

Number of samples with greater value of $\overline{S}^2$ statistic than the original at the four faculties

Faculty	Samples
Faculty of Animal Sciences	394
Faculty of Economic Sciences	0
Faculty of Arts	165
Faculty of Pedagogy	878

CONCLUSION

The open source R software environment is especially suitable to develop data analysing packages in large diversity. There are several packages to support the research work in the area of network analysis. The networksis package is suitable to examine stochastic network models.

We concluded that the study programmes offered by the faculties of Kaposvár University do not compete with each other. The analysis was done by separating the study programmes according to faculties. Further analysis might be needed to examine all the programmes pooled together.

REFERENCES

Admiraal, R., Handcock, M.S. (2008): networksis: A package to simulate bipartite graphs with fixed marginals through sequential importance sampling. J. Stat Softw. Vol. 24, Issue 8, <URL: http://www.jstatsoft.org/v24/i08/paper> [2009-06-30]

Csárdy, G.(2009): Package 'igraph' <URL: http://igraph.sourceforge.net/doc/R/igraph.pdf> [2009-06-30]

Handcock, M.S., Hunter, D.R., Butts, C.T., Goodreau, S.M., Morris, M. (2008): statnet: Software tools for the representation, visualization, analysis and simulation of network data. In: J Stat Softw. Vol. 24, Issue 1, <URL: http://www.jstatsoft.org/v24/i01/paper> [2009-06-30]

Ihaka, R., Gentleman, R . (1996): R: a language for data analysis and graphics. In: J-Comput. Graphic. Stat., 5: 299-314.

Roberts, A, Stone, L. (1990): Island-Sharing by Archipelago Species. In: Oecologia, 83: 560-567.

Solymosi, N. (2005): R <-…erre, erre…! (Bevezetés az R-nyelv és környezet használatába). <URL: http://cran.r-project.org/doc/contrib/Solymosi-Rjegyzet.pdf> [2009-06-30]

Corresponding author:

György KÖVÉR
Kaposvár University
Faculty of Economic Sciences
Department of Mathematics and Physics
H-7400 Kaposvár, Guba S. u. 40.
Tel.: +36-82-505-956
e-mail: kover.gyorgy@ke.hu