STATISTICAL PROPERTIES OF WEIGHTED MACROECONOMIC NETWORKS

Mircea_Gligor

National College “Roman Voda”, Roman-5550, Neamt, e-mail: mrgligor@yahoo.com

Abstract. *The properties of the weighted networks are investigated using some statistical physics tools, taking into account the statistical ensemble of the networks with fixed number of vertices. As application, the correlations between GDP/capita time series are investigated in various time windows, over the time interval 1993-2008. The target group of countries is the 27 EU members in 2008. The mean correlation coefficients are attached to the edges of a fully connected weighted network having the countries as nodes. Particularly, the concept of entropy, based on the probability of one particular realisation from the statistical ensemble, may yield some more information about the structure, stability and evolution of the EU country clusters.*

Keywords: *network, clusters, free energy, entropy*

PACS numbers: 89.65.Gh, 89.75.Fb, 05.45.Tp

1. INTRODUCTION

The study of nonequilibrium growing networks has become in the last years a well-defined field of research in theoretical physics and several large reviews are now available ([1] – [6]). In this section we aim at pointing out several important steps performed so far in two particular problems: the first one consist in the recent transition from the study of classical unweighted graphs to the weighted networks seen today as a better modeling of the most social, economic and ecological systems; the second one is related to the statistical mechanics methodology applied so far in the study of the unweighted networks, namely so-called “Hamiltonian approaches” of network statistical ensembles.

The goal of the present paper is to investigate the weighted fully connected network of the $N = 27$ countries forming the European Union in 2008 (EU-27). The ties between countries are supposed to be proportional to the degree of similitude of the macroeconomic fluctuations referring to the GDP/capita annual rates of growth between 1993 and 2008. The countries are abbreviated according to The Roots Web Surname List (RSL) [7] which uses 3 letters standardized abbreviations to designate countries and other regional locations. The World Bank database [8] is here used as data source.The common measure of 2-country fluctuation similarity is the (Pearson’s) correlation coefficient of the two time series describing the time evolution of the considered indicator. The correlation coefficient is calculated in a moving time window of $T = 5$ years size. The constant size time window is moved with 1 year time step until the full time interval 1993-2008 (containing $\Delta t = 16$ data points) is scanned. In this way a number of $\Delta t - T + 1 = 12$ correlation coefficients are obtained for each pair of countries.

The weight w_{ij} assigned to the network edge *i-j* are supposed to be equal to the average correlation coefficient of *i* and *j* countries. The correlation coefficients can only be averaged only by turning them into additive measures such as the coefficients of determination or Fisher *z*-scores. The averaging is effectuated in the both ways in Section 2 and the properties of the adjacency matrix eigensystem are comparatively analyzed. Using some arguments from the factor analysis, we find that for the cluster analysis goals, the averaging by means of coefficients of determination leads to better results than the averaging through *z*-values.

In Section 3, the actual EU-27 weighted network is seen as a particular realization from the statistical ensemble of networks having a fixed number of vertices. Mapping the weighted network into a multi-graph, the probability of this particular realization is calculated considering the links randomly attached between the $N = 27$ vertices. Some thermodynamic quantities, namely the entropy, free energy, mean energy/link and thermal susceptibility are defined following the standard tools of the classical statistical mechanics. The variation of these quantities is investigated, during a thinking process that consists in removing the countries one by one starting from the strongest connected ones and from the weakest connected ones respectively. Different paths of variation are found, and a sort of phase transition is identified from the variation of the thermal susceptibility. The economic meaning is straightforward, by observing that the transition point corresponds to the complete removal of some clusters of countries.

The conclusions and some further possible developments are given in Section 4.

2. AVERAGING THE CORRELATION COEFFICIENTS

An average of correlation coefficients in a number of samples does not represent an “average correlation” in all those samples. Because the value of the correlation coefficient is not a linear function of the magnitude of the relation between the variables, correlation coefficients cannot simply be averaged. In cases when one needs to average correlations, they first have to be converted into additive measures. The methods usually recommended in the statistics literature (e.g. [9]) are: (a) to square them and so to obtain the *coefficients of determination* which are additive, or, (b) to convert them into so-called *Fisher z* values, which are also additive.

The first method gives the average correlation coefficients of the form:

$$\hat{C}_{ij}^{(d)}(T) = \left[\frac{1}{\nu} \sum_{t=k}^{k+T} C_{ij}^{2}(t) \right]^{1/2} \quad , \ k = 0,\ 1,\ \ldots,\ \Delta t - T, \tag{1}$$

where Δt is the total number of points (the time span), T is the time window size used for the analysis, $\nu = \Delta t - T + 1$, and t is a discrete counter variable.

In order to apply the second method, the Fisher z-values are firstly calculated:

$$z_{ij} = \frac{C_{ij} - \mu}{\sigma}, \tag{2}$$

where μ and σ are the mean and the standard deviation of the C_{ij}'s distribution. The z-values are averaged as in Eq. (1) for the T-size time window moving over the Δt points of the dataset:

$$\hat{z}_{ij} = \frac{1}{\nu} \sum_{t=k}^{k+T} z_{ij}(t), \qquad k = 0,\ 1,\ \ldots,\ \Delta t - T\ . \tag{3}$$

At last, one comes back to the distribution having the mean μ and the standard deviation σ:

$$\hat{C}_{ij}^{(z)}(T) = \sigma \cdot \hat{z}_{ij}(T) + \mu \tag{4}$$

In the present approach, the considered time interval is between 1990 and 2005, i.e. Δt = 16 years, and the time window size is T = 5 years. The problem of the "optimal" choosing of the time window size is in close relation to the C_{ij}'s distribution, and it has been discussed elsewhere [10].

The clustering scheme of the N = 27 countries may be now constructed by the both ways. The correlation matrix eigensystems are analysed for $[C_{ij}^{(d)}]$ and $[C_{ij}^{(z)}]$ respectively.

Let us recall firstly that the eigenvalues can be interpreted as the proportion of variance explained by each canonical correlation relating two sets of variables. There will be as many eigenvalues as there are canonical correlations (roots), and each successive eigenvalue will be smaller than the last since each successive root will explain less and less of the data. In factor analysis, the eigenvectors of a correlation matrix correspond to factors, and eigenvalues to factor loadings. The observable random variables are modeled as linear combinations of the factors, plus the "error" terms.

Having a measure of how much variance each successive factor extracts, one can call the question of how many factors to retain. By its nature this is somehow an arbitrary decision. However, there are some guidelines that are commonly used [9], and that, in practice, seem to yield the best results. Firstly, we can retain only factors with eigenvalues greater than 1. In essence this is like saying that, unless a factor extracts at least as much as the equivalent of one original variable, one has to drop it. This criterion, firstly proposed by Kaiser [11], is probably the one most widely used. A graphical method is the "scree" test first proposed by Cattell [12]. Plotting the eigenvalues in a simple line plot, one has to find the place where the smooth decrease of eigenvalues appears to level off to the right of the plot. To the right of this point, presumably, one finds only "factorial scree" ("scree" is the geological term referring to the debris which collects on the lower part of a rocky slope). Both criteria have been studied in detail [13-15]. By generating random data based on a particular number of factors [13, 14], it was found that the first method (Kaiser criterion) sometimes retains too many factors, while the second technique (scree test) sometimes retains too few; However, both methods were found remarkably convergent when the number of common factors is not too large [15].

The above considerations explain why the first eigenvectors (i.e. the ones corresponding to the largest eigenvalues) are generally considered as carrying the useful information. The clustering scheme of the EU-27 countries is further constructed on the structure of these eigenvectors, pertaining to the correlation matrices $[C_{ij}^{(d)}]$ and $[C_{ij}^{(z)}]$.

In Table 1 one can see the first ten values from the two matrices eigenspectra. According to the above mentioned factor analysis criteria one has to retain two common factors when the averaging was done by means of the coefficients of determination and at least four when the averaging was done by means of the Fisher z-values.

Table 1 The first 10 eigenvalues of the correlation matrices constructed by averaging the coefficients of determination (the first row) and Fisher z-values (the second row)

Eval$[C_{ij}^{(d)}]$	**15.132**	**2.255**	1.159	1.077	0.912	0.719	0.663	0.603	0.505	0.428
Eval$[C_{ij}^{(z)}]$	**10.029**	**3.813**	**2.365**	1.940	1.412	1.206	0.976	0.860	0.676	0.397

In [10] the cluster structure of the EU-27 countries was done by token of the first two eigenvectors (***V*₁** and ***V*₂**) of $[C_{ij}^{(d)}]$. The countries can be partitioned into five groups, which we can call "Continental", "Scandinavian", "Mediteraneean", "Anglo" and "East-European". This partion is in good agreement with the results recently reported in the economic literature ([16-20]).

On the other hand one can see that using $[C_{ij}^{(z)}]$, the number of statistically significant eigenvalues is certainly greather than two. To have a more exact representation, in addition to the repreentation ($\boldsymbol{V}_1$ and $\boldsymbol{V}_2$), other two-eigenvector structures must be done: ($\boldsymbol{V}_1$ and $\boldsymbol{V}_3$) and ($\boldsymbol{V}_2$ and $\boldsymbol{V}_3$). Each important feature of the cluster structure that was done by token of the first two eigenvectors ($\boldsymbol{V}_1$ and

$\boldsymbol{V}_2$) of $[C_{ij}^{(d)}]$ can be recovered in at least one of the three representations ($\boldsymbol{V}_1$; $\boldsymbol{V}_2$), ($\boldsymbol{V}_2$; $\boldsymbol{V}_3$) and ($\boldsymbol{V}_1$; $\boldsymbol{V}_3$). The three representations are nothing else but orthogonal projections of the same *N*-points structure in the 3-dimensional eigenvector space. (Note that the full information would be derived in a 4-dimensional and 6-dimensional space respectively, in order to take into account *all* the significant eigenvalues in Table 1).

One can conclude that averaging the correlation coefficients by means of the *z*-values leads to a *more informative* but *harder comprehensible* clustering scheme, as compared to the one obtained by averaging by means of the coefficients of determination.

3. THE STATISTICAL TERMODYNMICS OF THE EU-27 WEIGHTED NETWORK

A. The statistical ensemble of networks with fixed number of vertices

Let us observe firstly that any weighted graph with $0 \leq w_{ij} \leq 1$ can be turn into a graph with positive integer weights and respectively into an unweighted multi-graph through a suitable multiplication of edge weights. For example, the weighted graph in Fig. 3a has been turned into the unweighted multi-graph in Fig. 3b by multiplying its weights by 10.

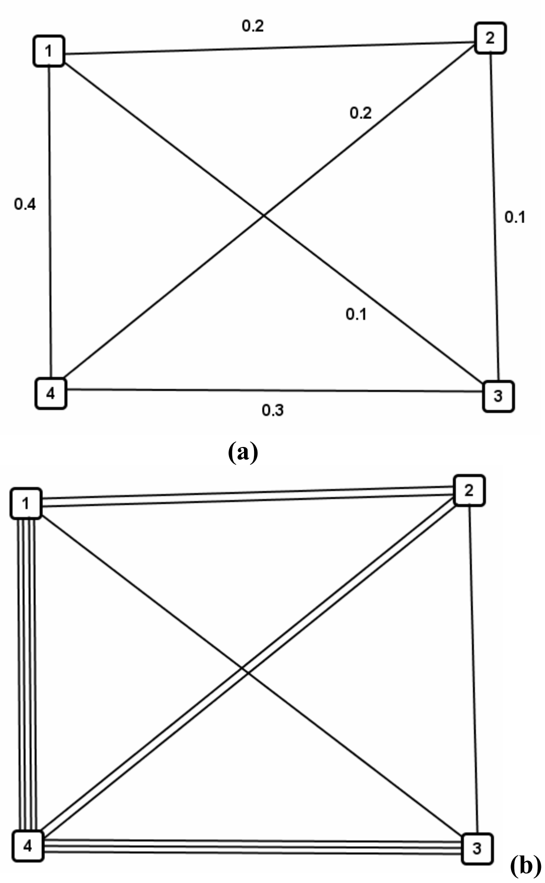

Fig. 3 The weighted graph (a) is turned into the multi-graph (b). The edge weights multiplied by 10 become positive integers and thus they can be mapped into unweighted multiple edges.

Now we can try to find the probability of having the weight w_{ij} assigned to the edge *i* - *j* on the hypothesis that in the isomorphic multi-graph the links are attached *randomly* between the *N* edges. For example, in Fig. 3b the probability of having three edges between the vertices (3) and (4) should be of the form: $p_{34} = C{\cdot}(1/6)^3$, where *C* is a positive constant accounting all the possible permutations of vertex labels, and (1/6) is the probability of having a link *somewhere* between a pair of vertices in the graph. If we have *N* vertices, the corresponding number of possible connections becomes: $\binom{N}{2} = N(N-1)/2$, and the probability of having w_{ij} simple edges between the vertices (*i*) and (*j*) is read:

$$p_{ij} = C\frac{1}{\binom{N}{2}^{w_{ij}}} = C\left(\frac{N(N-1)}{2}\right)^{-w_{ij}} .$$

Introducing the notation: $\Lambda = N(N-1)/2$, after the normalization:

$$\sum_{\substack{i,j\\ i>j}} p_{ij} = 1 ,$$

the above probability becomes:

$$p_{ij} = \frac{\Lambda^{-w_{ij}}}{\sum_{\substack{i,j\\ i>j}} \Lambda^{-w_{ij}}}$$

(9)

Finally, one can turn back to the initial network with $0 \leq w_{ij} \leq 1$; defining:

$$\beta = \ln \Lambda = \ln \frac{N(N-1)}{2} ,$$

(10)

Eq. (9) gets the more familiar "canonical" form:

$$p_{ij} = \frac{\exp(-\beta w_{ij})}{\sum_{\substack{i,j\\ i>j}} \exp(-\beta w_{ij})} .$$

(11)

Several remarks have to be done here.

(i) During the network "thinking" construction process, the links have been supposed to be *randomly* attached between the *N* vertices. Thus, the mechanism of network generating excludes any "preferential attachment". As a consequence, the probability in Eq. 11 has the typical form for an *exponentially growing network*, the term "growth" referring here at the increasing number of edges while the number of vertices has been kept constant.

(ii) The parameter β in Eq. 10 is not related to any temperature. Nonetheless, β can be seen as an *internal* parameter of the statistical ensemble of *N*-vertex networks, in the same way in which the temperature is for the canonical ensemble. Unlike the thermodynamic meaning, the changing of β does involve neither warming nor cooling process, but it simply means the shifting from a statistical ensemble to another one.

On the above assumptions, some basic thermodynamic quantities can be defined in correspondence to the classical statistical mechanics, as follows:

- The partition function:

$$Z = \sum_{\substack{i,j\\ i>j}} \exp(-\beta w_{ij}) \qquad (12)$$

- The entropy:

$$S = -\sum_{\substack{i,j\\ i>j}} p_{ij} \ln p_{ij} = -\sum_{\substack{i,j\\ i>j}} \frac{\exp(-\beta w_{ij})}{\sum\limits_{\substack{i,j\\ i>j}} \exp(-\beta w_{ij})} \ln \frac{\exp(-\beta w_{ij})}{\sum\limits_{\substack{i,j\\ i>j}} \exp(-\beta w_{ij})} \qquad (13)$$

- The free energy:

$$F = \frac{1}{\beta} \ln Z = \frac{1}{\beta} \ln \sum_{\substack{i,j\\ i>j}} \exp(-\beta w_{ij}) \qquad (14)$$

- The average energy / link:

$$< w >= \sum_{\substack{i,j\\ i>j}} p_{ij} w_{ij} = \sum_{\substack{i,j\\ i>j}} \frac{w_{ij} \exp(-\beta w_{ij})}{\sum\limits_{\substack{i,j\\ i>j}} \exp(-\beta w_{ij})} \qquad (15)$$

- The “thermal” susceptibility:

$$\Lambda\chi_T = \frac{d<w>}{d(1/\beta)} = -\beta^2 \frac{d<w>}{d\beta} = \beta^2\left[<w^2> - <w>^2\right] (16)$$

B. Deconstructing the EU-27 weighted network

In order to get some more information about the structure of the EU-27 weighted network we examine it during a thinking process of decomposition, which consist of removing the countries one by one, in decreasing and, respectively, increasing order of the overlapping coefficients displayed in Table 2. The overlapping coefficients (defined and calculated in Ref. [10] for all the EU-27 countries) are quantities able to measure to what extent a country is “connected” to the whole system.

Table 2 The country average overlapping index of each EU-27 country

SWE	**0.38**	**NLD**	**0.35**	**CYP**	**0.32**
DNK	**0.37**	**AUT**	**0.35**	**SVN**	**0.32**
GER	**0.37**	**FIN**	**0.35**	**CZE**	**0.31**
FRA	**0.37**	**POL**	**0.35**	**ROM**	**0.31**
HUN	**0.37**	**ESP**	**0.35**	**BGR**	**0.31**
SVK	**0.37**	**PRT**	**0.35**	**LTU**	**0.31**
BEL	**0.36**	**ITA**	**0.34**	**LVA**	**0.31**
IRL	**0.36**	**MLT**	**0.33**	**EST**	**0.30**
LUX	**0.36**	**GRC**	**0.33**	**GBR**	**0.29**

Keeping somehow the “thermodynamic” analogy, the quantities defined by Eqs. 12-16 are studied as functions of β, which is a measure of the number of remainder countries, and $(1 / \beta)$ that is a measure of the number of removed countries.

The results are plotted in Figs. 4-8. The entropy is found to have a power law dependence on β. (The country removing corresponds to decreasing values of β, so the deconstruction process in Fig. 4 and Fig. 5 can be followed by reading the x-axis from right to left). The two paths of deconstruction are not differentiated: the two plots collapses onto a single one.As opposite, the two paths of country removing appear as clearly differentiated in Fig. 5, where the free energy per link variation is represented in semi-log plot. The free energy per link is found to depend exponentially on β.

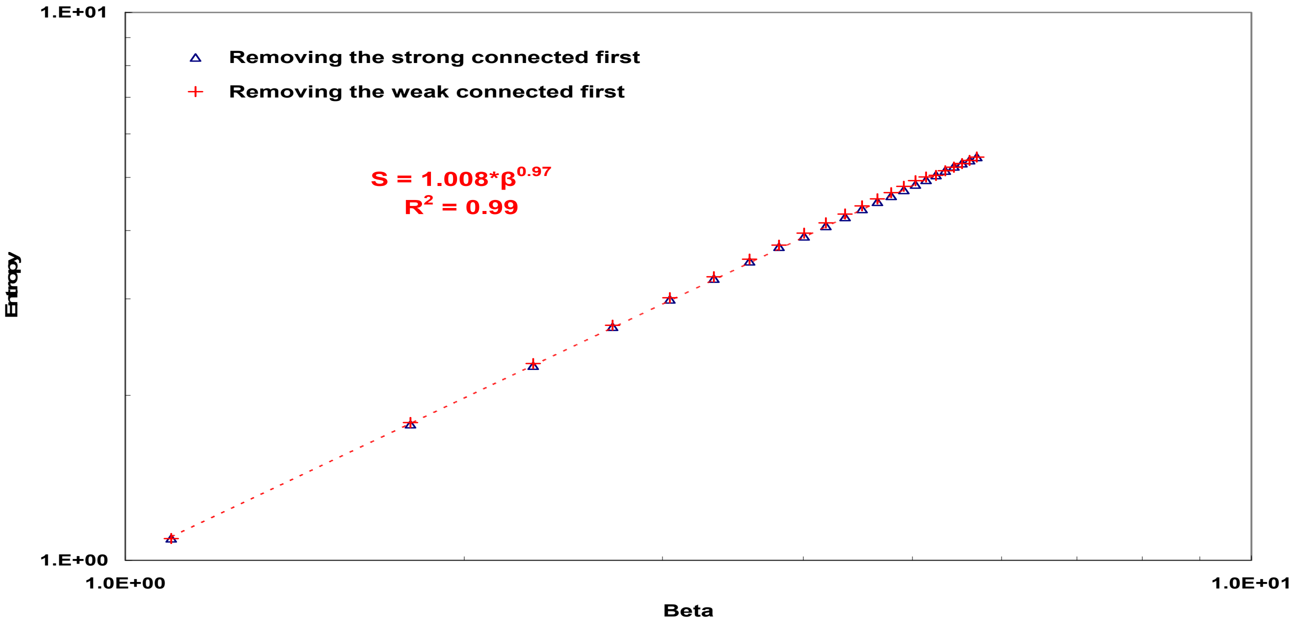

Fig. 4 The entropy variation in EU-27 country removing process, in log-log plot. The dashed line is the power law fit. R^2 is the square of Pearson's product moment correlation coefficient of fitting.

An interesting bi-fractal behavior displays the average energy per link when countries are removed in increasing order of connectness (Fig. 6). Such effect (if it exists) cannot be observed in the other path of deconstruction. One must stress here that removing the countries in decreasing order of connectness (i.e. starting with the strongest tied) generally involves large fluctuations at each step, thus this way of the deconstruction process can be seen as having a high level of noise.

The same noisy behavior displays the thermal susceptibility (Fig. 7). Nonetheless, as the statistical results are as more relevant as we have more elements in the system, we can restrict our analysis on the range of small values of $1/\ \beta$ (i.e. large values of N), where two local minima (one for each path of deconstruction) can be easily seen. They correspond to $(1/\ \beta)_1 = 0.25$ (i.e. 15 countries removed) when countries are removed in decreasing order of connectness, and to $(1/\ \beta)_2 = 0.19$ (i.e. 6 countries removed) when countries are removed in increasing order of connectness. From Table 2 one can easily find that this critical behavior corresponds to cluster separation (the East European cluster, GBR and the group MLT-GRC-CYP are separated in the first case; the East European cluster and GBR are separated in the second case).

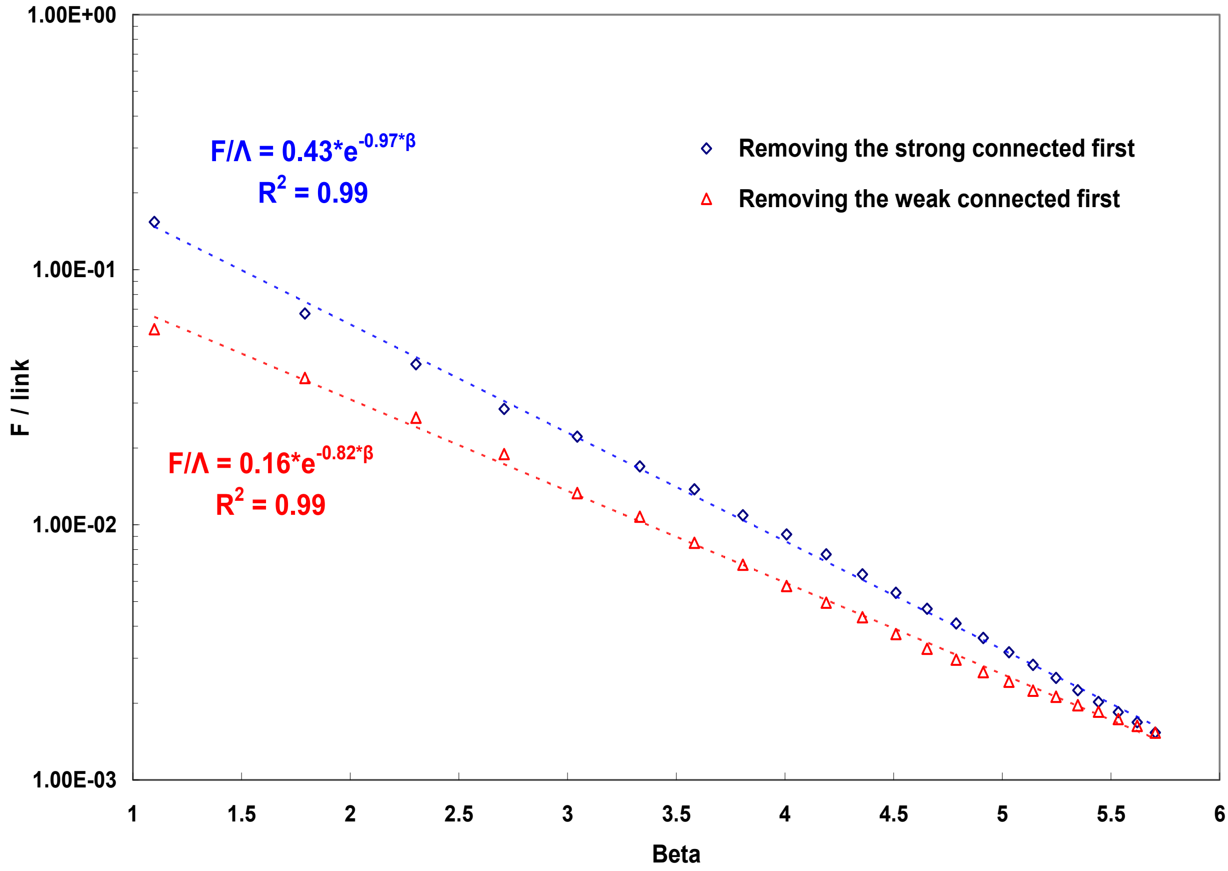

Fig. 5 The free energy per link variation in EU-27 country removing process in semi-log plot. The dashed lines are the exponential fits. R^2 is the square of Pearson's product moment correlation coefficient of fitting.

A well-established numerical method for analysis of critical points [21] consists in studying the temperature dependence of the fourth cumulant of $<w>$,

$$V_L = 1 - \frac{< w_{ij}^4 >}{3 < w_{ij}^2 >^2} \qquad (17)$$

This quantity is supposed to have a local minimum in the vicinity of the critical points, both for continuous and discontinuous phase transitions [22].

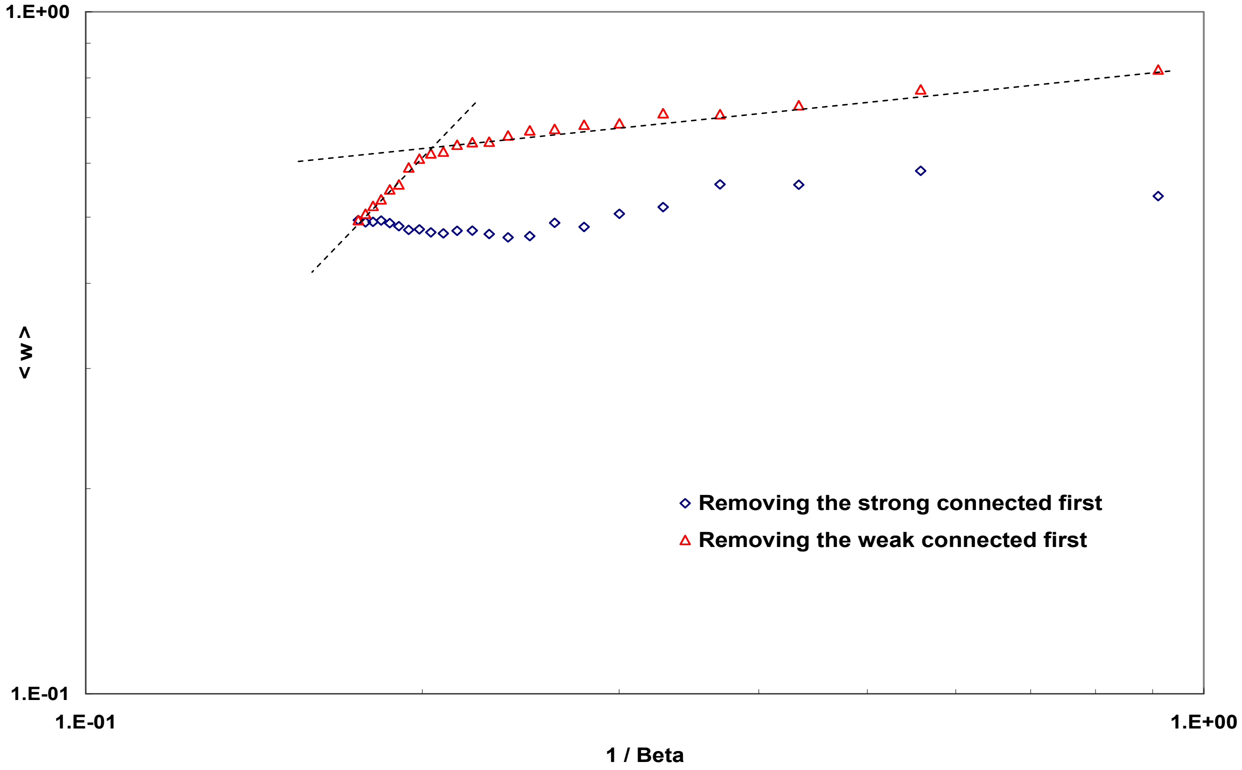

Fig. 6 The average energy per link variation in EU-27 country removing process in log-log plot. A bi-fractal behaviour is found when countries are removed in increasing order of connectness.

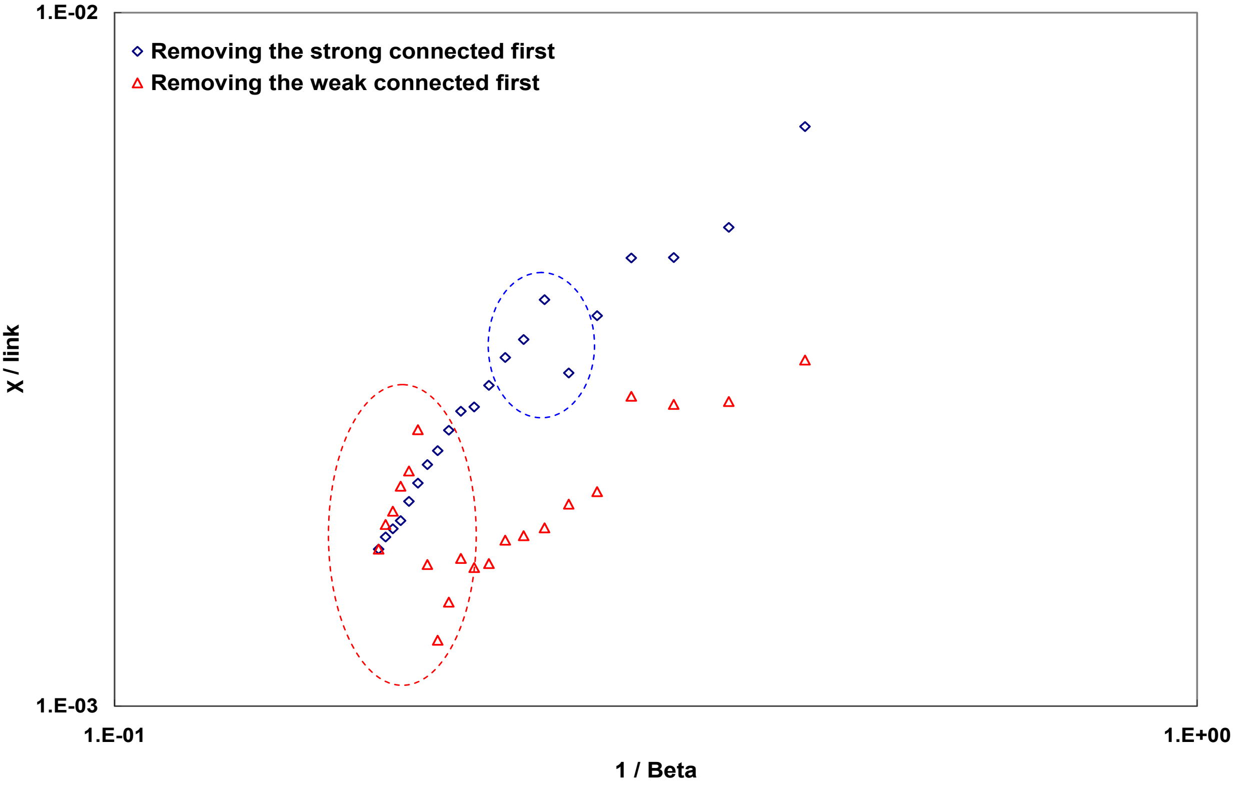

Fig. 7 The thermal susceptibility variation in EU-27 country removing process, in log-log plot. The first local maximum in each plot can be seen as critical point.

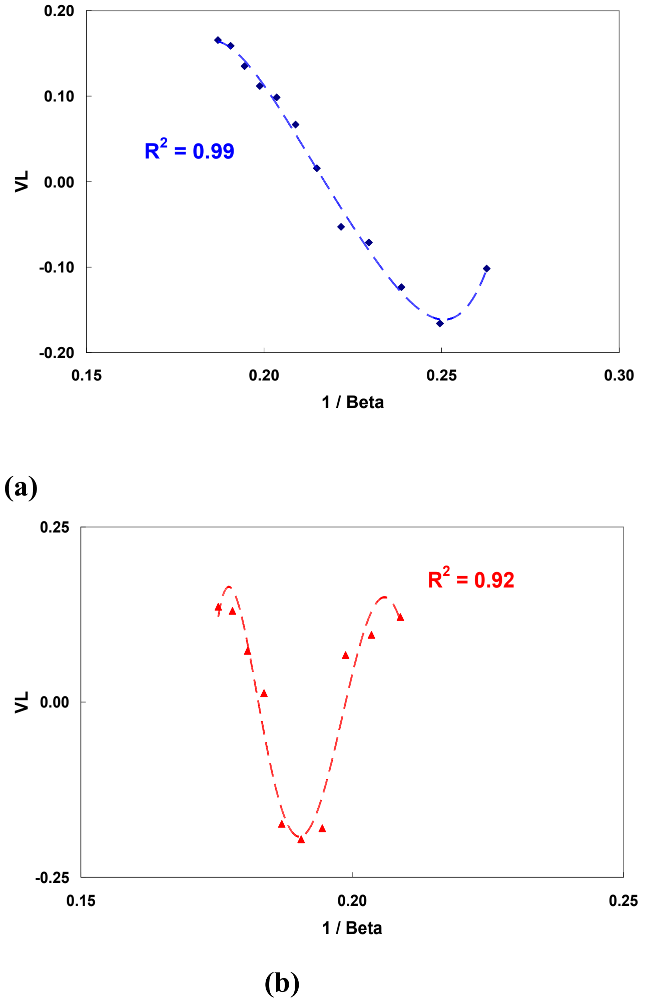

Fig. 8 The fourth central cumulant V_L variation in EU-27 country removing process, in the vicinity of the critical points. The minima correspond to the first local maxima of the thermal susceptibility found in Fig.7.

Given that the distribution mean and variance are changing from a statistical ensemble to another one, in the present approach the central moments are considered, thus Eq. (17) is read:

$$V_L = 1 - \frac{<(w_{ij} - <w_{ij}>)^4>}{3<(w_{ij} - <w_{ij}>)^2>^2} \tag{18}$$

This quantity is found to have two local minima in the vicinity of the two critical points above mentioned (Fig. 8).

4. CONCLUSION

Two recent topics in network analysis have been caught more an more interest in the last few years: the first one refers to extending the analytical investigations from the classical binary graphs to the weighted networks, which yield a more appropriate framework of modelling the real networks from communications, biology and economy; the second one, mostly restricted so far to the binary graphs, refers to construct a genuine statistical mechanics of networks, based on the fundamental notions of the field.

The goal of the present paper has been to join the two topics in a particular framework derived from mapping the macroeconomic time series into weighted graphs.The weights in the analyzed EU-27 weighted network represent some average correlation coefficients between the GDP/capita rates of growth, calculated for each pair of countries in a 5 years moving time window. The two ways of turning the correlation coefficients into additive measures have been comparatively analyzed in Section 2. It was proved that the averaging the correlation coefficients by means of the *z*-values as compared with the averaging by means of the coefficients of determination leads to a more informative but harder comprehensible clustering scheme.

The statistical ensemble of networks with fixed number of vertices was constructed and analyzed in Section 3. A probability has been assigned to each two-country connection by random attachment mechanism, and the corresponding partition function was built. The basic thermodynamic quantities, namely entropy, free energy, average energy per link and thermal susceptibility have been defined using the partition function. The variation of the thermodynamic quantities have been investigated during a thinking process of network deconstruction, which consist of removing the countries one by one, in decreasing and, respectively, increasing order of the overlapping coefficients. Some evidences for critical points have been found, the corresponding phase transitions being generated by removing compact clusters of countries from the system.

5. REFERENCES

[1] Albert R. & Barabási A.-L. (2002). *Statistical Mechanics of complex networks*. Review of Modern Physics 74, 47-97.

[2] Dorogovtsev S.N & Mendes J.F.F. (2003). *Evolution of Networks: From Biological Nets to the Internet and WWW*. Oxford University Press, Oxford.

[3] Pastor-Satorras R & Vespignani A. (2004). *Evolution and Structure of the Internet: A Statistical Physics Approach*. Cambridge University Press, Cambridge.

[4] Newman M.E.J. (2004). *Analysis of weighted networks*. Physical Review E 70, 056131.

[5] Ausloos, M. & Gligor, M (2008). *Cluster Expansion Method for Evolving Weighted Networks Having Vector-like Nodes*. Acta Physica Polonica A 114(3), 491-499.

[6] Gligor, M. & Ausloos, M (2008) *Clusters in weighted macroeconomic networks: the EU case*. European Physical Journal B 63, 533-539.

[7] http://helpdesk.rootsweb.com/codes/

[8] http://devdata.worldbank.org/query/default.htm

[9] Lewicki P. & Hill T. (2006). *Statistics. Methods and Applications*. StatSoft Inc. Tulsa, OK. Electronic version: http://www.statsoft.com/textbook/stathome.html

[10] Gligor, M. & Ausloos, M (2010). *Mapping macroeconomic time series into weighted networks*. Paper for the workshop EDEN 3, Piteşti, June, 15, 2010.

[11] Kaiser H.F. (1960). The application of electronic computers to factor analysis. Educational and Psychological Measurement 20, 141-151.

[12] Cattell R.B. (1966). *The scree test for the number of factor*. Multivariate Behavioral Research, **1**, 245-276.

[13] Linn R.L. (1968). A Monte Carlo approach to the number of factors problem". Psychometrika, 33, 37-71.

[14] Hakstian A.R., Rogers W.D., Cattell R.B. (1982). The behavior of numbers of factors rules with simulated data. Multivariate Behavioral Research, 17, 193-219.

[15] Browne M.W. (1968). A comparison of factor analytic techniques. Psychometrika 33, 267-334.

[16] Aaberge R., Bjorklund A., Jantti M., Palme M., Pedersen P.J., Smith N., Wennemo T. (2002). *Income Inequality and Income Mobility in the Scandinavian Countries Compared to the United States*. Review of Income and Wealth, 48, 443-469.

[17] Moran T.P. (2005). *Bootstrapping the LIS: Statistical Inference with the Gini Index and Patterns of Inequality in the Global North*. Paper for "International Conference in Memory of Two Eminent Social Scientists: C. Gini and M. O. Lorenz", Siena, Italy, 23-26 May, 2005.

[18] Durlauf S.N. & Quah D.T. (1999). *The new empirics of economic growth*. In *Handbook of Macroeconomics*. Elsevier, North Holland, 231–304.

[19] Mora T. (2005). *Evidencing European regional convergence clubs with optimal grouping criteria*. Applied Economics Letters 12, 937-940.

[20] Angelini E.C. & Farina F. (2005). *The size of redistribution in OECD countries: does it influence wave inequality?* Paper for "International Conference in Memory of Two Eminent Social Scientists: C. Gini and M. O. Lorenz", Siena, Italy, 23-26 May, 2005.

[21] Koza Z. & Ausloos M. (2007). *The Ising model in a Bak-Tang-Wiesenfeld sandpile*. Physica A 375, 199-211.

[22] Binder K. (1997). Applications of Monte Carlo methods to statistical physics. Rep. Prog. Phys. 60, 487.