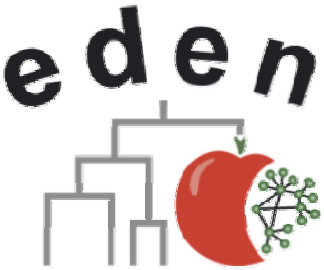

Project no. 043251

EDEN

ECOLOGICAL DIVERSITY AND EVOLUTIONARY NETWORKS

Instrument: Specific Targeted Project (STREP)

Thematic Priority: Integrating and strengthening the European Research Area
NEST Pathfinder initiative Tackling Complexity in Science

D2.2 (D5): Report on correlation methods

Due date of deliverable: Month 18

Actual submission date: Month 18

Start date of project: 1 January 2007 Duration: 36 months

Organisation name of lead contractor for this deliverable: **TKK/LCE**

Other contributors: **IMEDEA-UIB**

Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006)		
Dissemination Level		
PU	Public	**X**
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

TABLE OF CONTENTS

D2.2. Report on correlation methods

Network reconstruction based on correlations and distances

Summary

In this report we review recent advances in the analysis of fully connected networks with weighted links. The weight of a link can represent the similarity between a pair of nodes, e.g., a correlation, or their dissimilarity, e.g., a distance. While for the former several approaches have been developed to uncover the underlying topology, for the latter there isn't a standard procedure so far. An approach consists in mapping the weights of the links to a probability in the range [0,1]. Generalized measures defined in terms of the probabilities can then be applied. An alternative approach introduces a threshold allowing converting the weighted networks into directed networks with binary-valued adjacency matrices.

1. Introduction

The theory of complex networks, developed in the last decade, has proposed a set of techniques to analyse the topology of networks. A network is, in general, a representation of a system whose nodes are the elements of the system and the links represent interaction between pair of nodes. According to the properties of the nodes and links one can face different classes of networks. For example if there are two kinds of nodes, e.g. authors and papers, the network is bipartite; the links can be undirected, directed or weighted. Although there is a large literature dealing with undirected, and also directed, networks, the treatment of weighted networks is not so well developed. For unweighted complex networks, with binary adjacency matrices (i.e. matrices with elements taking one of two possible values, say 0 and 1, indicating the absence or presence or link), a set of local and global measures on the network has been defined, including the degree of a node, its average nearest-neighbor degree [Pastor-Satorras et al, 2001] and its clustering coefficient [Watts, Strogatz, 1998; Albert, Barabási 2002]. Defining these measures for weighted networks is more difficult and in some cases several unequivalent definitions have been given for the same concept. For instance a review of definitions of weighted clustering coefficients can be found in Ref. [Saramäki et al, 2007].

In the EDEN project we typically deal with fully-connected weighted networks, that is, for each pair of nodes there is a value which gives the similarity or dissimilarity between them. The question arises on how to extract relevant information from a network perspective. We can distinguish two cases: whether the weight reflects a similarity, i.e. larger values represent more similar, or dissimilarity, where a larger value indicates less similarity. An example of the first type of network is obtained when the weight is given by a correlation; in the remainder we will refer to these as *similarity networks.* Dissimilarity networks are obtained when the entries in the adjacency matrix measures distances between nodes; we will refer them as *distance networks*.

2. Similarity networks

Similarity networks can be obtained from time series. Some examples include:

- financial time series: where each node represents an asset, and the weight between two pair of assets is given by the time correlation of the asset price;
- brain activity: where the nodes represent an appropriate volume of the brain (being a neuron, a voxel or a brain area) and the weight between nodes is given by the time correlation of their activity recorded by different techniquess;

If $x_i(t)$ is the time evolution of the variable recorded at node *i*, the time correlation between nodes *I* and *j* is given by

$$C_{ij} = \frac{\langle x_i x_j \rangle - \langle x_i \rangle \langle x_j \rangle}{\sqrt{\langle x_i^2 \rangle \langle x_j^2 \rangle}},$$

where ⟨ ⟩ denotes temporal averages.

The magnitude $d_{ij} = \sqrt{2\left(1 - C_{ij}\right)}$ is a metric distance if C_{ij} is a time correlation.

Once a distance is introduced the typical analysis one can perform includes:

a) Minimum spanning tree: trees and graphs [see Fig. 1 for asset trees and graphs];
b) Spectral analysis, eigenvectors [see Fig.1].

Alternative methods aiming at generalizing measures defined in unweighted networks have been proposed and are introduced in Sections 3 and 4.

In the remainder we will consider weighted, fully connected networks. This means that for any pair of nodes (*i*, *j*) there is a weight w_{ij} obtained by a specified means. The matrix ***W*** is defined by the elements $\{w_{ij}\}$, where $i = \{1, \dots, N\}$ being *N* the number of nodes. In this report we will describe two approaches. The first one maps the adjacency matrix into a probability matrix ***P*** where the elements $\{p_{ij}\}$ are obtained in a convenient way from the matrix ***W***. The second approach converts the original

network into a directed unweighted network where each link is present or not depending on whether $w_{ij} > R$ for a pre-specified threshold *R*.

3. Mapping weighted, fully connected networks to ensemble *probability* networks

Similarity and distance matrices have been generated for example from time series, so that closely related series (i.e. with fluctuations highly correlated in time) have a large correlation coefficient among them. The first step is to find a continuous bijective map $M : R \rightarrow [0,1]$ from the real numbers to the interval between 0 and 1, which maps the weights $w_{ij} \in R$ to a quantity $p_{ij} \in R$. A simple example of such a map is a linear normalization of the weights:

$$p_{ij} = \frac{w_{ij} - \min(w_{ij})}{\max(w_{ij}) - \min(w_{ij})}.$$

This simple normalization maps min(w_{ij}) to zero. While this is often acceptable in the case of a correlation matrix, one should make a more sophisticated choice of map if there are many edges with weight min(w_{ij}). Similarly, if the network has negative weights as well as positive ones, the normalized modulus of the original weights might be a more appropriate choice. For the case in which the weights represent a distance instead of a correlation, analogous transformation can be easily defined. A more detailed discussion on the topic of map choice can be found in [Ahnert et al 2007].

The idea is to interpret the matrix ***P*** with entries $\{p_{ij}\}$ as a matrix of probabilities, an ensemble of edges or *ensemble network*. Thus, just as any binary square matrix can be understood as an unweighted network and any real square matrix corresponds to a weighted network, any square matrix with entries between 0 and 1 corresponds to an ensemble network. If we sample each edge of the ensemble network exactly once, we obtain an unweighted network which we term a realization of the ensemble network. In particular, p_{ij} is the probability that the edge between nodes *i* and *j* exists. These concepts are valid both for directed networks, with any $p_{ij} \in [0, 1]$, and undirected networks, for which $p_{ij} = p_{ji}$, so that the matrix is symmetric. In a real-world weighted network, the original weights can represent almost any physical quantity, such as the

strength of collaboration between two scientists, the number of passengers traveling between two countries, or genetic distance. By mapping these weights to probabilities we rid ourselves of the interpretational burden of these weights, whilst retaining all the topological information they contain. It should be noted that in many cases the interpretation of weights as probabilities also makes intuitive physical sense. Whenever the weights in a network represent a magnitude of flow, this can be interpreted directly in terms of the probability that a transfer occurs during a given unit of time. Examples include traffic and transport networks as well as communication networks, where we have units (passengers, money, signals) which form an edge, through their transfer, with a probability proportional to the flow rate. We hope that the same ideas would apply also to networks representing gene flow, one of the targets of EDEN.

All measures on unweighted networks can be written as functions of the entries a_{ij} of an adjacency matrix $\boldsymbol{A}$. In fact, generally they can be written as a polynomial of these entries, or a simple ratio of such polynomials. Note that, for an unweighted network, $a_{ij} = (a_{ij})^m$ for all positive integers $m > 0$, so that these polynomials are of first order only. Consider a general first-order polynomial, which can be written fully expanded as:

$$f(\boldsymbol{A}) = \sum_{q=0}^{2^{N^2}} C_q \prod_{j,k=0}^{N} a_{jk}^{b(q)_{jk}} , \qquad (1)$$

where N is the number of nodes, the C_q are real coefficients and the $b(q)_{jk}$ are a set of Boolean matrices specifying which adjacency matrix entries appear in each term of the polynomial. The probability P_q that $\prod_{j,k=0}^{N} a_{jk}^{b(q)_{jk}} = 1$ in a given realization $\boldsymbol{A}$ is simply $P(q) = \prod_{j,k=0}^{N} P_{jk}^{b(q)_{jk}}$. Thus, due to the linearity of the polynomial, the average $\bar{f}(\mathbf{P})$ of f over the ensemble network realizations is:

$$\bar{f}(\boldsymbol{P}) = \sum_{q=0}^{2^{N^2}} C_q \prod_{j,k=0}^{N} p_{jk}^{b(q)_{jk}} = f(\boldsymbol{P}) \ . \qquad (\mathbf{2})$$

This means that the value of a polynomial function f of the entries of an unweighted network $\boldsymbol{A}$, averaged over the realizations of a given ensemble network $\boldsymbol{P}$ is equal to the value of the polynomial of the ensemble network adjacency matrix itself.

Average degree. The degree k_i of a given node *i* in an unweighted network with adjacency matrix elements a_{ij} is the number of its neighbors, and is written as $k_i = \sum_j a_{ij}$. In a weighted network with elements w_{ij} the corresponding quantity has been termed the strength of the node *i*, denoted as s_i, which consists of the sum of the weights: $s_i = \sum_j w_{ij}$. In an ensemble network, the corresponding sum over the edges attached to a particular node gives the average degree of node *i* across realizations, denoted as $\overline{k_i}$ and given by $\overline{k_i} = \sum_j p_{ij}$.

It is important to note that while the strength of a node in a weighted network may have meaning in the context of the network, $\overline{k_i}$ has a universal meaning, regardless of the original meaning of the weights.

Clustering coefficient. The clustering coefficient of a node *i*, which has been defined [Watts, Strogatz 1998] as:

$$c_i = \frac{\sum_{j,k} a_{ij} a_{jk} a_{kl}}{k(k-1)} = \frac{\sum_{j,k} a_{ij} a_{jk} a_{kl}}{\sum_{j,k} a_{ij} a_{ik}} \quad , \qquad (3)$$

where $k \neq j \neq i \neq k$ in the sums. This corresponds to the number of triangles in the network which include node *i*, divided by the number of pairs of bonds including *i*, which represent potential triangles. Using the ensemble approach with its normalized weights this generalizes straightforwardly to:

$$c_i^e = \frac{\sum_{j,k} p_{ij} p_{jk} p_{kl}}{\sum_{j,k} p_{ij} p_{ik}} \quad , \qquad (4)$$

which can be read as the average number of triangles divided by the average number of bond pairs. In modified form, this clustering coefficient has appeared in the very recent literature [Zhang, Horvath, 2005] but without connection to a general approach to the construction of weighted network measures based on a general mapping from weights to probabilities. Note that c_i^e is not the average of c_i over the ensemble. For a detailed discussion of this subtlety, see [Ahnert et al, 2007].

All measures constructed with the ensemble approach are only functions of the normalized weights p_{ij} , not of the elements of an unweighted adjacency matrix a_{ij} or of

the degree *k*. This distinguishes the ensemble measures from measures proposed for weighted networks in the literature, such as the weighted clustering coefficient c_i^w:

$$c_i^w = \frac{1}{k(k-1)/2} \sum_{j,k} \frac{(w_{ij} + w_{ik})}{2} a_{ij} a_{ik} a_{jk} \quad , \qquad (5)$$

and the weighted average nearest-neighbor degree $k^w{}_{nn,i}$:

$$k_{nn,i}^w = \frac{1}{s_i} \sum_{j=1}^{N} a_{ij} w_{ij} k_j \quad . \qquad (6)$$

Both are defined in Ref. [Barrat et al, 2004], and eq. (5) is the most frequently cited definition of a weighted clustering coefficient in the literature. Due to their construction, these measures cannot be used for the analysis of fully connected weighted networks, as $k_{nn,i}^w = 1$ and $c_i^w = 1$ for all nodes *i* in such networks. Fully connected weighted networks form an important class of complex networks, for example in the form of the (virtually fully-connected) EU air travel network [Ahnert et al, 2007], functional brain networks [Eguíluz et al, 2005], genetic networks [Rozenfeld et al 2007]. Furthermore, any matrix of similarities or distances between a number of objects can be treated as a fully connected weighted network, and thus can be analyzed using the ensemble approach, but not with approaches such as eq. (5) and (6), which are "mixed" in the sense that they make use of both the unweighted and weighted adjacency matrix entries.

Note that the absolute values of the ensemble clustering coefficient have limited meaning, as they are dependent on the map *M*. It is their relative values which carry the information, and these are largely independent of the choice of map *M*, as long as it is bijective.

4. Thresholding

A complementary approach when dealing with a large number of nodes consist in thresholding the weights: for a given threshold value *R* a network, its adjacency matrix, can be constructed such that if the weight is larger than the threshold then a link

connecting the pair of nodes is present otherwise the link is absent. This idea can be applied both to similarity and to distance networks (Figure 2). Following this approach a weighted network can be transformed in a set of unweighted networks. This approach has been followed to analyze brain functional networks [Eguíluz et al, 2005]. From time series of brain activity obtained using fMRI, undirected networks have been obtained describing brain functional networks for healthy humans performing simple cognitive tasks. The topological analysis includes scale-free degree distributions, small-world property (small path length and large clustering), and degree-degree correlations [Figure 3].

The question is now which is the convenient threshold to be considered, as the properties of the network change depending on the threshold value *R*: for small values of *R*, the reconstructed network is fully connected, while if the value of *R* is large enough, the network is composed by isolated nodes. For intermediate values it has been found that the network can display a percolation transition where the network displays a large connected component. Percolation properties have been analyzed in some detail in cell-phone networks [Onnela et al 2007] finding that the properties of the network depend on whether the percolation is analyses removing the strongest link or the weakest links (Figure 4).

References

R. Albert, A.-L. Barabási, *Statistical mechanics of complex networks*, Rev. Mod. Phys. 74, 47 (2002).

S. E. Ahnert et al, Ensemble approach to the analysis of weighted networks, Phys. Rev. E 76, 016101 (2007).

S. E. Ahnert et al, *Applying weighted network measures to microarray distance matrices,* J. Phys A: Math. Theor. 41, 224011 (2008).

A. Barrat et al, *The architecture of complex weighted networks*, Proc. Natl. Acad. Sci. USA 101, 3747 (2004).

V.M. Eguíluz et al, *Scale-free brain functional networks*, Phys. Rev. Lett. **92**, 018102 (2005).

J.P. Onnela et al, *Structure and tie strengths in mobile communication networks,* Proc. Natl. Acad. Sci. USA 104, 7332-7336 (2007).

T. Heimo et al, *Spectral and network methods in the analysis of correlation matrices of stock returns*, Physica A 387, 147 (2007).

R. Pastor-Satorras et al, *Dynamical and Correlation Properties of the Internet*, Phys. Rev. Lett. 87, 258701 (2001).

A.F. Rozenfeld et al, *Genetic diversity and networks of genetic similarity*, J. of the Royal Soc. Interface 4, 1093–1102 (2007)

J. Saramäki et al, *Generalizations of the clustering coefficient to weighted complex networks*, Phys. Rev. E 75, 027105 (2007).

D.J. Watts, S.H. Strogatz, *Collective dynamics of 'small-world' networks*, Nature 393, 440 (1998).

B. Zhang, S. Horvath, *A General Framework for Weighted Gene Co-Expression Network Analysis,* Stat. Appl. Gen. Mol. Biol. 4, 17 (2005).

Figure 1. Asset graph for the NYSE. Clusters corresponding to the eigenvectors of the correlation matrix, identified by the clique percolation method, are denoted by the shaded background. Reproduced from Ref. [Heimo et al, 2007]

Figure 2. Methodology used to extract functional networks from the signals. The correlation matrix is calculated and then used to define the network among the highest correlated nodes. Top four images represent snapshots of activity and the three traces correspond to selected voxels from visual (V1), motor (M1) and posterio-parietal (PP) cortices. Reproduced from Ref. [Eguíluz et al 2005]

Figure 3. Degree distribution for three values of the correlation threshold. The inset depicts the degree distribution for an equivalent randomly connected network. Reproduced from Ref. [Eguíluz et al 2005]

Figure 4. The stability of the mobile communication network to link removal. The control parameter f denotes the fraction of removed links. (A and C) These graphs correspond to the case in which the links are removed on the basis of their strengths (w_{ij} removal). (B and D) These graphs correspond to the case in which the links were removed on the basis of their overlap (O_{ij} removal). The black curves correspond to removing first the high-strength (or high O_{ij}) links, moving toward the weaker ones, whereas the red curves represent the opposite, starting with the low-strength (or low O_{ij}) ties and moving toward the stronger ones. (A and B) The relative size of the largest component $R_{GC}(f) = N_{GC}(f)/N_{GC}(f=0)$ indicates that the removal of the low w_{ij} or O_{ij} links leads to a breakdown of the network, whereas the removal of the high w_{ij} or O_{ij} links leads only to the network's gradual shrinkage. (A Inset) Shown is the blowup of the high w_{ij} region, indicating that when the low w_{ij} ties are removed first, the red curve goes to zero at a finite f value. (C and D) According to percolation theory, $\tilde{S} = \sum_{s<s_{max}} n_s s^2/N$ diverges for $N \to \infty$ as we approach the critical threshold f_c, where the network falls apart. If we start link removal from links with low w_{ij} (C) or O_{ij} (D) values, we observe a clear signature of divergence. In contrast, if we start with high w_{ij} (C) or O_{ij} (D) links, there the divergence is absent. Finite size scaling shows that the small local maximum seen in D at $f \cong 0.95$ does not correspond to a real phase transition. Reproduced from Ref. [Onnela et al, 2007]

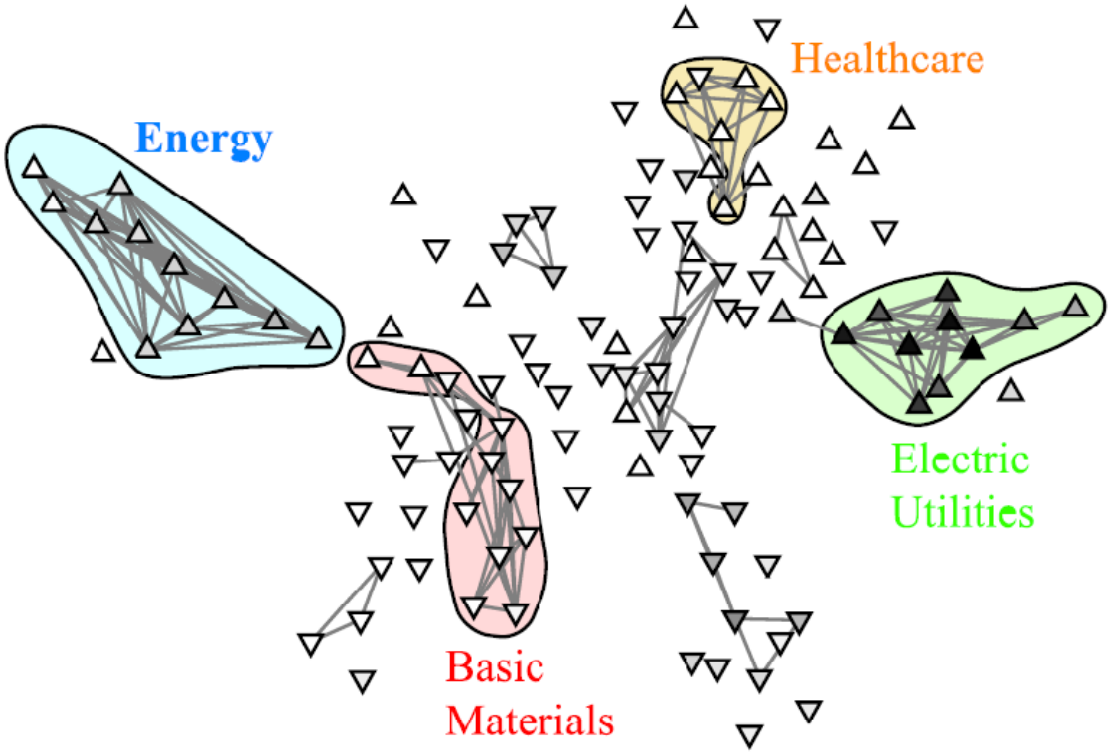

Figure 1

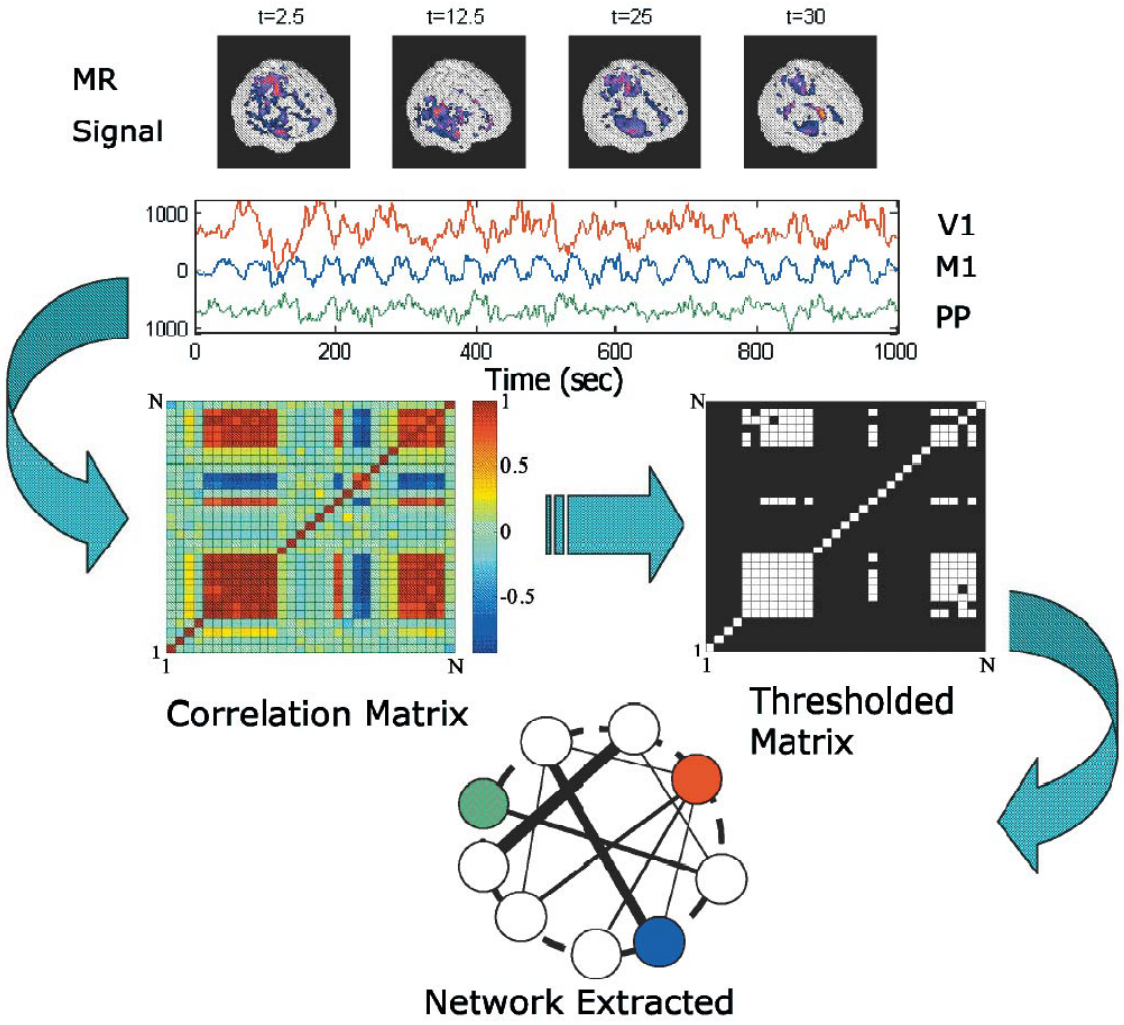
t=2.5
t=12.5
t=25
t=30
MR
Signal
1000
0
1000
0
200
400
600
800
1000
Time (sec)
V1
M1
PP
N
1
1
N
1
0.5
0
-0.5
Correlation Matrix
Thresholded
Matrix
Network Extracted

Figure 2

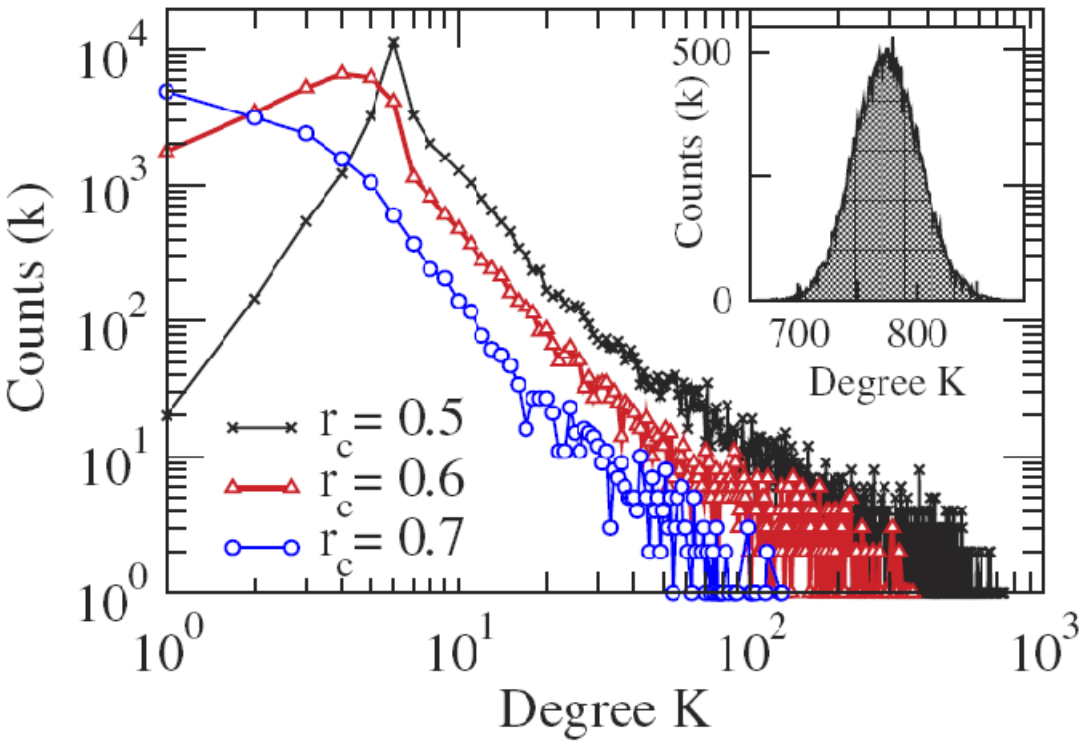

Figure 3

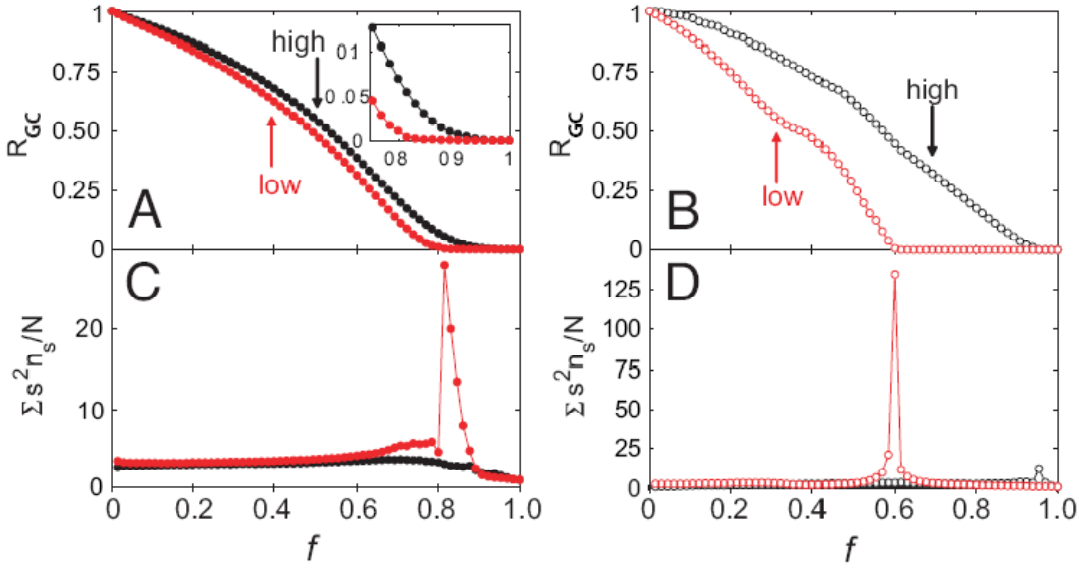
A
B
C
D
high
low
high
low
R_{GC}
$\Sigma s^2 n_s/N$
f
0
0.2
0.4
0.6
0.8
1.0
0.25
0.50
0.75
10
20
25
50
75
100
125
0.05
08
09

Figure 4