Ensemble approach to the analysis of weighted networks

S. E. Ahnert,[1] D. Garlaschelli,[2] T. M. A. Fink,[3] and G. Caldarelli[4,5]

[1]*Institut Curie, CNRS UMR 144, 26 rue d'Ulm, 75248 Paris, France*
[2]*Dipartimento di Fisica, Universitá di Siena, Via Roma 56, 53100 Siena, Italy*
[3]*Systems Biology and CNRS UMR 144, Institut Curie, 26 rue d'Ulm, 75248 Paris, France*
[4]*INFM-CNR Istituto dei Sistemi Complessi and Dipartimento di Fisica, Universitá di Roma "La Sapienza" Piazzale Moro 2, 00185 Roma, Italy*
[5]*Centro Studi e Museo della Fisica Enrico Fermi, Compendio Viminale, 00185 Roma, Italy*

We present an approach to the analysis of weighted networks, by providing a straightforward generalization of any network measure defined on unweighted networks, such as the average degree of the nearest neighbors, the clustering coefficient, the "betweenness," the distance between two nodes, and the diameter of a network. All these measures are well established for unweighted networks but have hitherto proven difficult to define for weighted networks. Our approach is based on the translation of a weighted network into an ensemble of edges. Further introducing this approach we demonstrate its advantages by applying the clustering coefficient constructed in this way to two real-world weighted networks.

INTRODUCTION

Weighted complex networks appear in many different contexts, for example, when studying transport and traffic [1,2], in the form of trade or communication networks, financial networks [3], and collaboration networks [4], to name a few. In addition, high-throughput technology has generated large amounts of biological data which can be interpreted in terms of weighted networks, such as networks of genetic regulation and transcription [5] and protein interaction [6]. While such networks can now be generated relatively easy, the extraction of meaningful physical or biological information from these networks is a much more challenging task. For unweighted complex networks—in which the entries of the adjacency matrix are restricted to zero and one—a set of local and global measures on the network has been defined [7], including the degree of a node, its average nearest-neighbor degree [8], and its clustering coefficient [9]. Further measures include the distance between two nodes, the related diameter of the network and the betweenness [10] of an edge or node. While the definition of such measures for unweighted networks is relatively straightforward, defining these measures for weighted networks is more difficult and has been the subject of recent research [2,5,11–13].

Here we introduce a new approach to this problem which allows for a straightforward generalization of any measure defined on an unweighted network to weighted networks. In addition we explicitly construct weighted versions of the clustering coefficient, the average degree of neighbors, the distance between two nodes and the diameter of the network. We compare this newly constructed clustering coefficient to a weighted clustering coefficient in the literature and to a version used in unweighted networks. The data sets we use for this comparison are aviation passenger data within the European Union (EU), which constitutes an almost fully connected network, and the network formed by neighboring letters in the English language.

ENSEMBLE NETWORKS

The basis of our approach is to find a continuous bijective map $M:\mathbb{R} \rightarrow [0,1]$ from the real numbers to the interval between 0 and 1, which maps the weights $w_{ij} \in \mathbb{R}$ to a quantity $p_{ij} \in [0,1]$. A simple example of such a map is a linear normalization of the weights

$$p_{ij} = \frac{w_{ij} - \min(w_{ij})}{\max(w_{ij}) - \min(w_{ij})}. \qquad (1)$$

This simple normalization maps $\min(w_{ij})$ to zero. This is often acceptable in the case of a distance matrix, but if there are many edges with weight $\min(w_{ij})$, one should introduce a parameter $\epsilon \ll 1$, such that

$$p_{ij} = \frac{w_{ij} - \min(w_{ij}) + \epsilon}{\max(w_{ij}) - \min(w_{ij}) + \epsilon}. \qquad (2)$$

Many other more sophisticated maps are imaginable and the final choice of map depends on the properties of the physical system underlying the network and the resulting distribution of weights. Appropriately chosen maps can deal with all variants of weighted networks including those with negative weights, and with differing interpretations of $w_{ij}=0$ as meaning "no edge" or as a physical weight. We will return to the topic of map choice below.

The ideas we introduce in this paper are based on an interpretation of the matrix **P** with entries $\{p_{ij}\}$ as a matrix of probabilities. These probabilities can be interpreted as an ensemble of edges, or more concisely, an 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. Note that, while some specific weighted networks discussed in the literature have probabilities as their weights [5,14], a general framework for the analysis of weighted networks, based on the transformation of weights to probabilities, has to our knowledge not been proposed. In a real-world weighted network, the original weights can represent almost any physical quantity, such as the strength of a collaboration between two scientists, or the number of passengers traveling between two countries. This is why we use a map M to translate the original weights into probabilities. Doing so does not destroy any of the topological information contained in the weights and connections, but allows us to analyze this information in the unifying framework which the probabilities provide. Furthermore, in many cases of real-world weighted networks, the transformation of weights to probabilities has a physical meaning. Examples include flow networks of traffic and transport, communications networks as well as collaboration networks. In all these, the interactions between nodes involve the transfer of a discrete unit (e.g., passengers, currency, or data packets) over a given period of time. Thus the weight, representing the number of units transferred, is directly related to the probability of observing the transfer of a unit at a given point in time.

In the framework of ensemble networks any existing measure on unweighted networks can be turned into an equivalent measure on weighted networks. A suitable choice of map M depends on the distribution of weights. For example, in both real-world networks which we analyze in this paper the original weights w_{ij} take values across several orders of magnitude, so that we chose the p_{ij} to be the normalized logarithms of the original weights, rather than the normalized weights themselves.

POLYNOMIALS OF ADJACENCY MATRIX ENTRIES

All measures on unweighted networks can be written as functions of the entries a_{ij} of an adjacency matrix **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

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

where N is the number of nodes, C_q are real coefficients, and $b(q)_{jk}$ 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 **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

$$\bar{f}(\mathbf{P}) = \sum_{q=0}^{2^{N^2}} C_q \prod_{j,k=0}^{N} p_{jk}^{b(q)_{jk}} = f(\mathbf{P}). \qquad (3)$$

This means that the value of a polynomial function f of the entries of an unweighted network **A**, averaged over the realizations of a given ensemble network **P**, is equal to the value of the polynomial of the ensemble network adjacency matrix itself. We will illustrate the power of this result in the following sections.

CONSTRUCTING THE MEASURES

Our approach allows for the construction of weighted network measures from their unweighted counterparts. As almost all existing unweighted measures are designed for undirected networks, the measures we construct in the remainder of this paper are also undirected. In general, however, our method is equally well suited to the transformation of any measure for directed, unweighted networks into one for directed and weighted networks. 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 $k_i=\Sigma_j a_{ij}$. 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: $s_i=\Sigma_j w_{ij}$. 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 $\bar{k}_i$ and given by $\bar{k}_i=\Sigma_j p_{ij}$.

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, $\bar{k}_i$ has a universal meaning, regardless of the original meaning of the weights. Now consider the total number of edges n in a network—also referred to as its size—given by $n=\Sigma_{ij} a_{ij}$ in the directed case and half this value in the undirected case where $a_{ij}=a_{ji}$. Replacing a_{ij} by p_{ij} again gives us the average size $\bar{n}$ of the realizations of the ensemble network, which is simply $\bar{n}=\Sigma_{ij} p_{ij}$ (or half this value for the undirected case).

A more complex measure in unweighted networks is the average degree of the nearest neighbors k_i^{NN}, which is the number of neighbors of i's neighbors, divided by the number of neighbors of i [8]:

$$k_i^{NN} = \frac{\sum_j k_j}{k_i} = \frac{\sum_{j,k} a_{ij}a_{jk}}{\sum_j a_{ij}},$$

where $j \neq i$ in the sums. By rewriting k_i^{NN} solely in terms of the a_{ij}, this generalizes to ensemble networks in a very straightforward manner:

$$k_i^{NN,e} = \frac{\sum_{j,k} p_{ij}p_{jk}}{\sum_j p_{ij}}.$$

This measure $k^{NN,e}$ is simply a ratio of averages: the average number of neighbors of i's neighbors over the average number of i's neighbors.

For unweighted networks the clustering coefficient of a node i has been defined [9] as

$$c_i = \frac{\sum\limits_{j,k} a_{ij}a_{jk}a_{ik}}{k(k-1)/2} = \frac{\sum\limits_{j,k} a_{ij}a_{jk}a_{ik}}{\sum\limits_{j,k} a_{ij}a_{ik}}, \tag{4}$$

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

$$c_i^e = \frac{\sum\limits_{j,k} p_{ij}p_{jk}p_{ik}}{\sum\limits_{j,k} p_{ij}p_{ik}} \tag{5}$$

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 [5] 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 $k^{NN,e}$ and c_i^e are *not* the averages of k_i^{NN} and c_i over the ensemble. We will address this subtlety below.

As an example of the power of Eq. (3), consider the distance d_{ij} (i.e., the shortest path) between two nodes i and j in an unweighted N-node network, represented entirely as a function of adjacency matrix entries

$$d_{ij}(\mathbf{A}) = a_{ij} + (1 - a_{ij})\sum_{m=1}^{N}(m+1)\alpha_{ij}^{(m)}(\mathbf{A})\beta_{ij}^{(m)}(\mathbf{A}),$$

where $\alpha_{ij}^{(m)}(\mathbf{A}) = \Pi_{q=1}^{m-1}[1 - \beta_{ij}^{(q)}]$, with

$$\beta_{ij}^{(m)}(\mathbf{A}) = 1 - \prod_{k_1,\ldots,k_m}\left(1 - a_{ik_1}a_{k_m j}\prod_{l=1}^{m-1}a_{k_l k_{l+1}}\right),$$

where all Π without a range are equal to one. As d_{ij} is a first-order polynomial in a_{ij}—the elements of the adjacency matrix $\mathbf{A}$—we know immediately from Eq. (3) that the average distance in the ensemble network will be $\bar{d}_{ij}(\mathbf{P}) = d_{ij}(\mathbf{P})$. Thus we have defined a distance measure on weighted networks without having to define a pairwise distance function of the edge weights [such as, for example, $d_{ij} = (w_{ij})^{-1}$ [4]].

Similarly, the *diameter* of an unweighted network, defined as the maximum distance $D(\mathbf{A}) = \max d_{ij}(\mathbf{A})$ between two nodes out of all pairs of nodes i, j can be written as a first-order polynomial

$$D(\mathbf{A}) = \prod_{p,q}a_{pq} + \sum_{m=1}^{N}(m+1)\zeta^{(m)}(\mathbf{A})\xi^{(m)}(\mathbf{A}),$$

where $\zeta^{(m)}(\mathbf{A}) = \Pi_{q=1}^{m-1}[1 - \xi^{(q)}(\mathbf{A})]$ and $\xi^{(m)}(\mathbf{A}) = \Pi_{i,j}\beta_{ij}^{(m)}(\mathbf{A})$. This expression allows us to straightforwardly calculate the average diameter $\bar{D}(\mathbf{P}) = D(\mathbf{P})$ of the ensemble network.

Another measure, the betweenness [10] of a node i or an edge (i, j), is the number of different shortest paths in the network which run through i or (i, j)s in the network. Similar to measures such as the distance and diameter, the betweenness can also be generalized to the weighted case by simply replacing the a_{ij} by p_{ij}. As the expressions in terms of adjacency matrix entries are rather involved, we do not give them here explicitly.

Some measures on unweighted networks, such as the average neighbor degree k_{NN} and the clustering coefficient c_i are *ratios* of two adjacency matrix polynomials f and g, which in general can be written as $h(\mathbf{A}) = f(\mathbf{A})/g(\mathbf{A})$. Now we can *define* the quantity $h^e(\mathbf{P}) \equiv \bar{f}(\mathbf{P})/\bar{g}(\mathbf{P}) = h(\mathbf{P})$. But, as was pointed out above, this quantity is no longer an average of $h(\mathbf{A})$ itself [which would be denoted $\bar{h}(\mathbf{P})$]. This gives us two distinct classes of measures: The first contains measures which can be written in polynomial form, and for which the ensemble version gives the average across realizations. These measures represent countable, integer quantities of the network, such as the number of neighbors, the number of triangles, the length of the shortest path between i and j, and so on. The second class are measures which are ratios of polynomials, such as the average nearest-neighbor degree or the clustering coefficient. The ensemble network version of these measures gives the ratio of the averages.

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:

$$c_i^w = \frac{1}{s_i(k_i - 1)}\sum_{j,k}\frac{(w_{ij} + w_{ik})}{2}a_{ij}a_{ik}a_{jk} \tag{6}$$

and the weighted average nearest-neighbor degree $k_{NN,i}^w$:

$$k_{NN,i}^w = \frac{1}{s_i}\sum_{j=1}^{N}a_{ij}w_{ij}k_j. \tag{7}$$

Both are defined in Ref. [2]. 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 which we analyze in this paper. Furthermore *any* matrix of similarities or distances between a number of objects—such as, for instance, microarray data series in biological experiments—can be treated as a fully connected weighted network, and thus can be analyzed using the ensemble approach, but not with approaches such as Eqs. (6) and (7), which are "mixed" in the sense that they make use of both the unweighted and weighted adjacency matrix entries.

ANALYZING REAL-WORLD WEIGHTED NETWORKS

In the following we demonstrate some of the advantages which the ensemble approach has over unweighted network measures, as well as over mixed weighted network measures. We do this by applying the ensemble clustering coefficient of
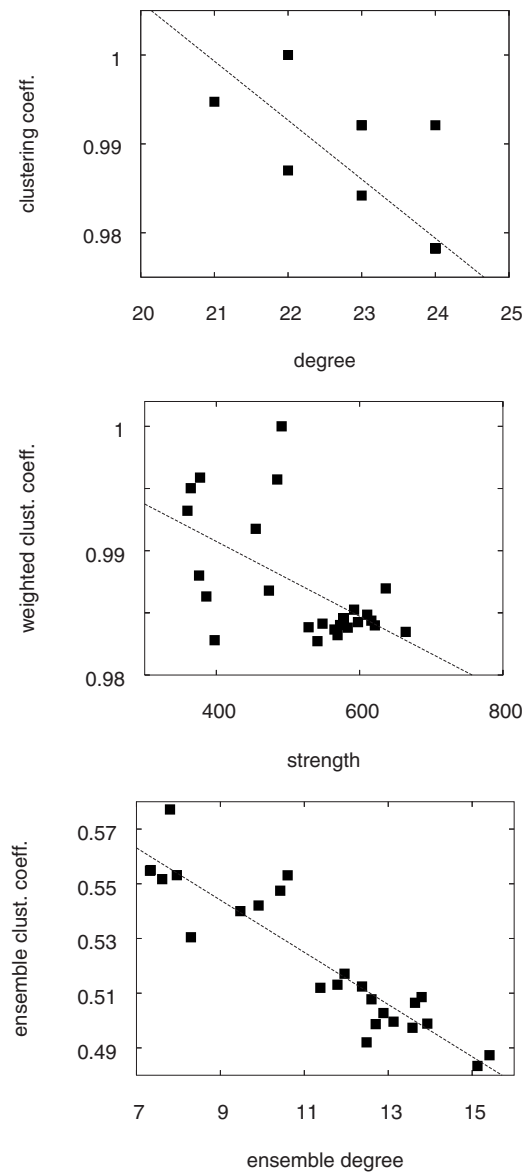
FIG. 1. Analysis of the network of air travel passengers within the 25 member states of the EU. This network is almost fully connected. Top: Unweighted clustering coefficient versus degree. All 25 data points are projected onto 7 locations, as a result of the information loss due to discarding the weights, and because the network is almost fully connected. Middle: Clustering coefficient as proposed in the literature [2] versus strength. This "mixed" clustering coefficient is a function of unweighted and weighted quantities. No clear relationship is evident, again because the network is almost fully connected. Bottom: Ensemble clustering coefficient versus ensemble degree. Unlike the other two approaches, those derived using the ensemble quantities exhibit a clear negative linear relationship. The lines are lines of best fit. Note that the absolute scale of the ensemble clustering coefficient c_i^e depends on the choice of the map M from weights to probabilities, which makes the relative values of c_i^e more important than the absolute ones.

Eq. (5) to two real-world networks. The first is the network of passengers traveling by air within the EU during 2004 [16]. The second is a network of letters in the English language, where the weight of the edge between two letters is
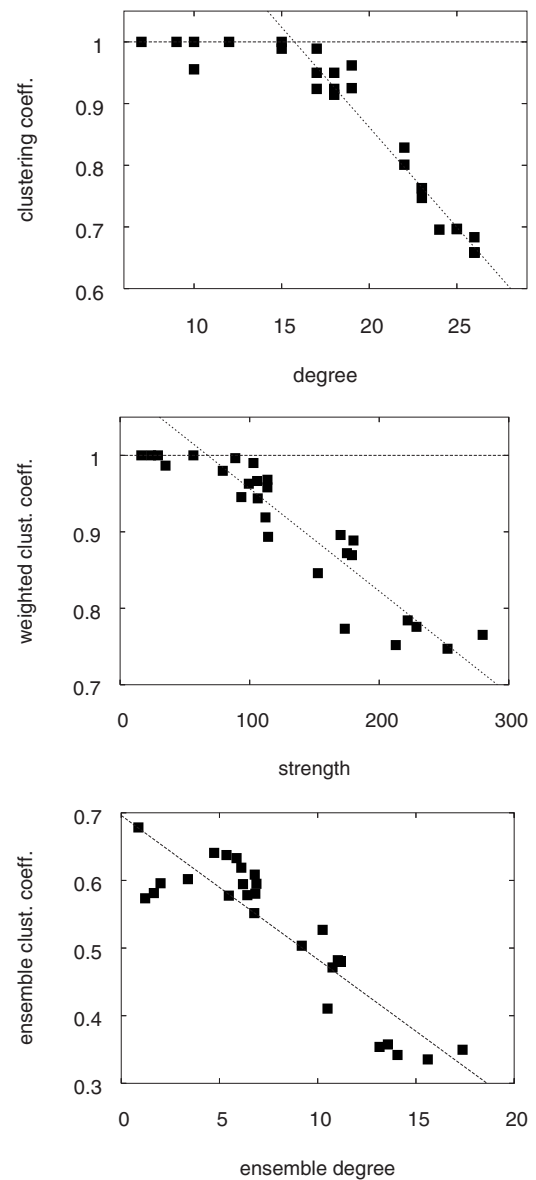

FIG. 2. Analysis of the weighted network formed between the 26 letters of the alphabet and the space between words [15]. Top: Unweighted clustering coefficient versus degree. Middle: Clustering coefficient as proposed in the literature [2] versus strength. This "mixed" clustering coefficient is a function of unweighted and weighted quantities. Bottom: Ensemble clustering coefficient versus ensemble degree. The ensemble approach makes use of all information contained in the weights, while the two others lose some of the information, as is shown by the plateau which both exhibit on the left side of the plots. The diagonal lines are lines of best fit for data points below the plateau. Note that the absolute scale of the ensemble clustering coefficient c_i^e depends on the choice of the map M from weights to probabilities, which makes the relative values of c_i^e more important than the absolute ones.

determined by the frequency at which they appear next to each other in the English language [15]. Both networks include edges which lead from a node to itself. The network of letters has 485 edges between 27 nodes (the alphabet and space), and therefore is 62.8% connected, while the EU net-

work with 607 connections between the 25 member states of the EU is almost fully (97.1%) connected.

In Fig. 1 we show the analysis of the EU air travel network using three different clustering coefficients: the unweighted clustering coefficient c_i of Eq. (4) [9], the mixed weighted clustering coefficient c_i^w of Eq. (6) with weighted and unweighted components from the literature [2] and the clustering coefficient c_i^e of Eq. (5) derived from the ensemble approach. These quantities are plotted against the degree k (in the unweighted case), the strength s_i for the mixed approach, and the ensemble degree $\bar{k}_i$ in for the ensemble approach. As this network is almost fully (97.1%) connected, the difficulty of the unweighted and mixed approaches becomes apparent: For the unweighted case, the 25 nodes of the network are mapped to just 7 points, representing the information lost by dropping the weights. In the mixed case, little can be deduced about the relationship between the clustering coefficient c_i^w and the strength s_i. The ensemble approach on the other hand reveals a clear negative linear relationship between the ensemble clustering coefficient c_i^e of Eq. (5) and the ensemble degree $\bar{k}_i$. Note that the absolute values of the ensemble clustering coefficient do not mean very much, 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. Countries with a large number of air passengers traveling in and out have a high ensemble degree $\bar{k}_i$ but also a low ensemble clustering coefficient c_i^e, as the many countries to which they are strongly connected are mostly not well-connected themselves. Thus these nodes with low c_i^e are surrounded by few triangles in any given ensemble realization, but many potential triangles in the form of pairs of edges. The inverse argument is true for nodes with a low ensemble degree $\bar{k}_i$, as any two neighbors of such a node are more likely to be strongly connected. For example, the two countries at the bottom right of the plot (high $\bar{k}_i$, low c_i^e) are the UK and Germany, while the top left corner (low $\bar{k}_i$, high c_i^e) contains Lithuania, Estonia, and Slovakia.

In Fig. 2 we show the analysis of the letter network using the same three clustering coefficients. As the letter network is less than two-thirds (62.8%) connected, the unweighted and mixed approaches do not encounter the difficulties associated with fully connected networks. However, if there are clusters in the network which are fully connected on a local scale—such that all neighbors of a given node are fully connected—these approaches again cannot differentiate any further between such nodes. In both the unweighted and mixed cases the letters Q, Z, J, and V are affected, as these letters only have few neighbors, which are fully connected among themselves, making the unweighted and mixed clustering coefficients equal to one. In Fig. 2 these four letters are represented by the four data points on the plateau which appears in the plots for the unweighted and mixed measures. No information, however, is lost with the ensemble approach, which again shows a clear negative linear relationship between ensemble clustering coefficient and ensemble degree. As before, the implication of this is that nodes with many strong connections—in this case the vowels A, E, I, O, and U, which are located at the bottom right of the plots in Fig. 2—have neighbors which are weakly connected among each other. These are the consonants, which are mostly located in the top left corner (low $\bar{k}_i$, high c_i^e).

CONCLUSION

We have introduced a general approach for the construction of measures on weighted networks, by introducing the concept of an ensemble network, in which every edge has a probability p_{ij} of existing. By transforming a weighted network into an ensemble network, any of the numerous measures which have been defined for unweighted networks can be straightforwardly generalized to weighted networks. Using the clustering coefficient constructed in this way as an example we demonstrate that these measures on weighted networks can reveal the additional topological information given in the weights, in particular for fully connected networks.

[1] A. de Montis, M. Barthelemy, A. Chessa, and A. Vespignani, e-print arXiv:physics/0507106.

[2] A. Barrat, M. Barthélemy, R. Pastor-Sarras, and A. Vespignani, Proc. Natl. Acad. Sci. U.S.A. **101**, 3747 (2004).

[3] J.-P. Onnela, A. Chakraborti, K. Kaski, J. Kertész, and A. Kanto, Phys. Rev. E **68**, 056110 (2003).

[4] M. E. J. Newman, Phys. Rev. E **64**, 016131 (2001).

[5] B. Zhang and S. Horvath, Stat. Appl. Gen. Mol. Bio. **4** (1), Article 17 (2005).

[6] L. Salwinski *et al.*, Nucleic Acids Res. **32**, D449 (2004).

[7] R. Albert and A.-L. Barabási, Rev. Mod. Phys. **74**, 47 (2002).

[8] R. Pastor-Satorras, A. Vázquez, and A. Vespignani, Phys. Rev. Lett. **87**, 258701 (2001).

[9] D. J. Watts and S. H. Strogatz, Nature (London) **393**, 440 (1998).

[10] L. C. Freeman, Sociometry **40**, 35 (1977).

[11] J.-P. Onnela, J. Saramäki, J. Kertész, and K. Kaski, Phys. Rev. E **71**, 065103(R) (2005).

[12] M. E. J. Newman, Phys. Rev. E **70**, 056131 (2004).

[13] J. Saramäki, M. Kivelä, J.-P. Onnela, K. Kaski, and J. Kertész, e-print arXiv:cond-mat/0608670.

[14] P. Grindrod, Phys. Rev. E **66**, 066702 (2002).

[15] M. S. Mayzner and M. E. Tresselt, Psychol. Med. Monogr. Suppl. **1**, 13 (1965).

[16] These data can be downloaded from the *Eurostat* website http://epp.eurostat.cec.eu.int