Applying weighted network measures to microarray distance matrices

Applying weighted network measures to microarray distance matrices

S. E. Ahnert

Theory of Condensed Matter Group, Cavendish Laboratory, JJ Thomson Avenue, Cambridge CB3 0HE, UK

D. Garlaschelli

Dipartimento di Fisica, Universitá di Siena, Via Roma 56, 53100 Siena, Italy

T. M. A. Fink

Institut Curie, CNRS UMR 144, 26 rue d'Ulm, 75248 Paris, France

G. Caldarelli

INFM-CNR Istituto dei Sistemi Complessi and Dipartimento di Fisica Universitá di Roma "La Sapienza" Piazzale Moro 2, 00185 Roma, Italy, and Centro Studi e Museo della Fisica Enrico Fermi, Compendio Viminale, 00185 Roma, Italy

Abstract.
In recent work we presented a new approach to the analysis of weighted networks, by providing a straightforward generalization of any network measure defined on unweighted networks. This approach is based on the translation of a weighted network into an ensemble of edges, and is particularly suited to the analysis of fully connected weighted networks. Here we apply our method to several such networks including distance matrices, and show that the clustering coefficient, constructed by using the ensemble approach, provides meaningful insights into the systems studied. In the particular case of two data sets from microarray experiments the clustering coefficient identifies a number of biologically significant genes, outperforming existing identification approaches.

The rise of information technology and the internet, as well as the more recent advent of high-throughput technologies in biology make it easier to obtain large amounts of data on complex networks. Increasingly this also includes data on weighted complex networks, which now appear in many different guises: Transport and traffic [1, 2], trade or communication networks, financial networks [3], and collaboration networks [4], to name a few. In biology, genetic regulation and transcription [5] and protein interaction [6] have been studied in this context. However, the extraction of meaningful physical or biological information from these networks is a difficult task. For unweighted complex networks, with binary adjacency matrices, a set of local and global measures on the network has been defined [7], including the *degree* of a node, its *average nearest-neighbour degree* [8] and its *clustering coefficient* [9]. Defining these measures for weighted networks is more difficult and has been the subject of recent research [2, 5, 10, 11]. A review of definitions of weighted clustering coefficients can be found in [12].

In a recent paper [13] we introduced a new approach to this problem which allows for a straightforward generalization of any measure defined on an unweighted network to weighted networks. Here we apply the clustering coefficient defined in this way to distance matrices, which are fully connected weighted networks. The distance matrices are generated from microarray expression series, so that closely related series (by some chosen similarity measure) will be separated by a short distance, which in the network picture translates into an edge with a large weight.

The basis of our approach is to find a continuous bijective map $M : \mathbb{R} \to [0, 1]$ 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:

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

This simple normalization maps $\min(w_{ij})$ to zero. While this is often acceptable in the case of a distance 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. A more detailed discussion on the topic of map choice can be found in [13].

The ideas we introduce in [13] are based on an interpretation of the matrix $\mathbf{P}$ 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. 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. 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.

All measures on unweighted networks can be written as functions of the entries a_{ij} of an adjacency matrix $\mathbf{A}$. 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, 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 $\mathbf{A}$ 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:

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

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

The degree k_i of a given node i in an unweighted network with adjacency matrix elements a_{ij} is the number of its neighbours, 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 $\bar{k_i}$ and given by $\bar{k_i} = \sum_j p_{ij}$.

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, $\bar{k_i}$ has a universal meaning, regardless of the original meaning of the weights.

As a more complex example, consider the *clustering coefficient* of a node i, which has been defined [9] as:

$$c_i = \frac{\sum_{j,k} a_{ij} a_{jk} a_{ik}}{k(k-1)/2} = \frac{\sum_{j,k} a_{ij} a_{jk} a_{ik}}{\sum_{j,k} a_{ij} a_{ik}} \tag{3}$$

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:

$$c_i^e = \frac{\sum_{j,k} p_{ij} p_{jk} p_{ik}}{\sum_{j,k} p_{ij} p_{ik}} \tag{4}$$

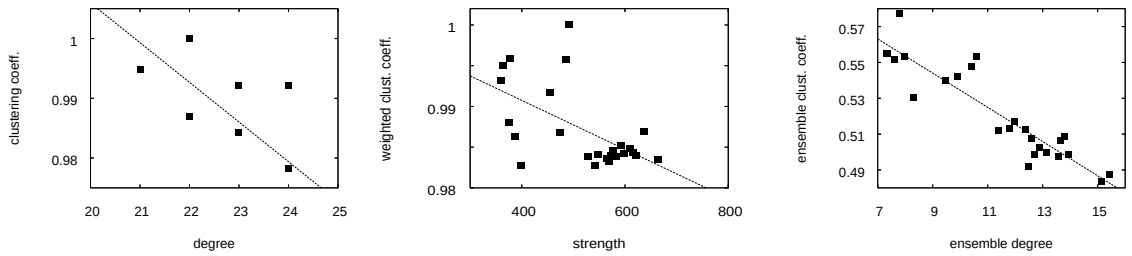

Figure 1. Example of the advantages of the ensemble clustering coefficient, as shown in our earlier work [13]: The network of air travel passengers within the 25 member states of the EU[15] is almost fully connected. LEFT: Unweighted clustering coefficient versus degree.All 25 data points are projected onto 7 locations, as a result of the information loss due to discarding the weights, and because the network is almost fully connected. CENTER: 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. RIGHT: 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.

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 c_i^e is *not* the average of c_i over the ensemble. For a detailed discussion of this subtlety, see [13].

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:

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

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

$$k_{nn,i}^w = \frac{1}{s_i} \sum_{j=1}^{N} a_{ij}w_{ij}k_j \tag{6}$$

Both are defined in [2], 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 which we analyze in [13] (see Fig. 1). 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

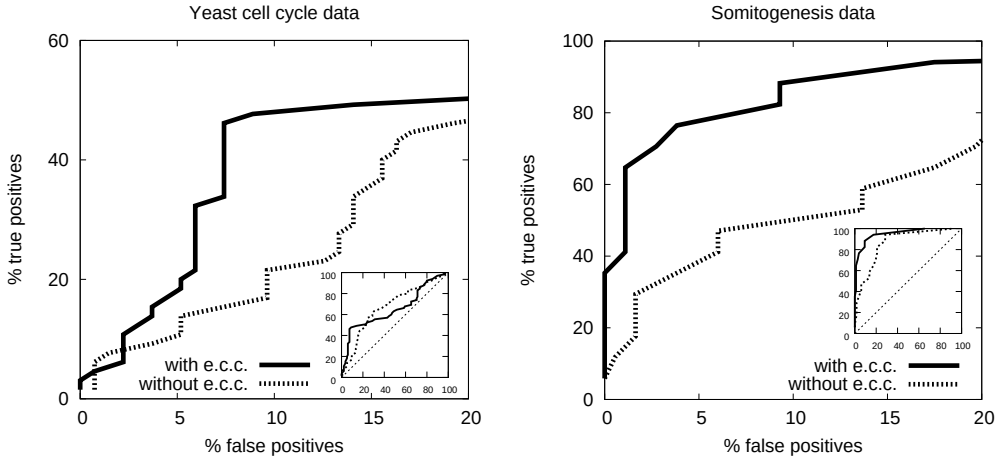

Figure 2. Receiver-operating characteristic (ROC) diagrams for the yeast cell cycle (LEFT) and somitogenesis (RIGHT) datasets, showing the positions of known biologically significant genes in a ranking of 200 genes in the rankings generated (a) using the ensemble clustering coefficient (solid) and (b) using the original pattern-finding approach (dotted) which was used to select the 200 genes in the first place. In both cases the ensemble clustering coefficient moves biologically significant genes to the top of the ranking.

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.

Microarrays are one of the most successful high-throughput technologies in biology, providing a snapshot of gene expression levels for all of the thousands of genes in the genome of a given organism simultaneously. A microarray consists of a large number of microscopic spots on a slide (typically made of glass or silicon), which each contain copies of a different short DNA sequence (or *oligonucleotide*) unique to a particular gene. Furthermore, the sequence copies in each spot are attached to a flourescent marker. If a given gene is expressed in the tissue sample to be examined, many copies of this gene will be present in the form of messenger RNA (mRNA), which in turn will bind to the sequences on the microarray, causing flourescence of the spot. The flourescence of the array of spots is captured by a camera and then read out using a computer.

A series of microarray measurements gives an expression profile for each gene over space or time, telling us where and when a given gene is 'switched on'. These sets of data series are subjected to detailed analysis, and distance matrices between these series, (often calculated using Pearson correlation) typically form an integral part of such an analysis.

Here we calculate the ensemble clustering coefficient for distance matrices derived from two entirely different microarray data sets. The first data set

consists of microarray data from an experiment studying the formation of vertebra (somitogenesis) in mice [16], from which a list of 200 genes was compiled using an existing pattern detection approach [17]. This approach is designed to detect biologically significant genes by finding expression profiles which deviate from randomness. The second data set is the well-known dataset of yeast cell cycle microarray experiments in yeast [18]. Here too the 200 strongest patterns were selected using the same approach.

It should be noted that microarray datasets are notoriously noisy and pre-filtering of data based on purely mathematical measures is essential and in fact present in almost any microarray study. Our selection method based on pattern detection is mathematically rigorous and makes no prior assumptions about the nature of the pattern.

In each of the two datasets the 200 genes are ranked by the amount of pattern they contain (and thus by their supposed biological significance). Yet the fully connected weighted network which corresponds to a distance matrix between these 200 genes contains none of this information. Therefore, when we calculate the ensemble clustering coefficient for a distance matrix of 200 genes, we can use the pattern-detection approach as a benchmark comparison for the performance of the clustering coefficient in finding biologically significant genes.

For both the mouse somitogenesis and yeast cell cycle datasets we compare our predictions to lists of known biologically significant genes. In the case of mouse somitogenesis these are 17 genes associated with the Wnt and Notch pathways, listed in [16], and in the case of yeast cell cycle there are 65 genes which can be found in two lists of experimentally verified yeast cell cycle genes [19, 20].

The distance measure chosen to generate the distance matrix is the algorithmic compression of one expression series due to another [17]. As can be seen in Fig. 2, the ranking generated by using the clustering coefficient clearly outperforms the pattern-ranking for both datasets. In the case of the mouse somitogenesis dataset, 11 (64%) of the 17 genes known to play a role in somitogenesis are located in the top 13 places (top 6%) of the ranking. Similarly, in the yeast cell cycle dataset, 31 (48%) of 65 known genes occupy places in the top 43 (top 21%). Compared to this, the conventional pattern-finding approach fares less well, with 6 (35%) in the top 13 (somitogenesis) and 23 (35%) in the top 43 (yeast). The conclusion is that in both datasets the ensemble clustering coefficient appears to move biologically significant genes to the top of the ranking.

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. As we have shown in this paper, our approach is particularly suited for the analysis of distance matrices. We demonstrate this by calculating the ensemble clustering coefficient for the distance matrices between microarray data series which successfully identifies many known biologically significant genes. These results are an indication that the application of complex networks methods to the rather separate field of distance matrix analysis is likely to yield valuable insights.

[1] A. de Montis, M. Barthelemy, A. Chessa, and A. Vespignani, Environment and Planning B - Planning & Design **34**, 905 (2007)

[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. Kertsz, and A. Kanto Phys. Rev. E **68**, 056110 (2003)

[4] M. E. J. Newman, Phys. Rev. E **64**, 016131 and 016132 (2001)
[5] B. Zhang and S. Horvath, Stat. Appl. Gen. Mol. Bio. **4**(1): Article 17 (2005)
[6] L. Salwinski *et al.*, Nuc. Ac. Res. **32** D449 (2004)
[7] R. Albert and A.-L. Barabási, Rev. Mod. Phys. **74**, 47 (2002)
[8] R. Pastor-Sartorras, A. Vázquez, and Alessandro Vespignani, Phys. Rev. Lett. **87**, 258701 (2001)
[9] D. J. Watts and S. H. Strogatz, Nature **393**, 440 (1998)
[10] J.-P. Onnela, J. Saramäki, J. Kertèsz, and K. Kaski, Phys. Rev. E **71**, 065103(R) (2005)
[11] M. E. J. Newman, Phys. Rev. E **70**, 056131 (2004)
[12] J. Saramaki, M. Kivela, J.-P. Onnela, K. Kaski, and J. Kertesz, Phys. Rev. E **75**, 027105 (2007)
[13] S. E. Ahnert, D. Garlaschelli, T. M. A. Fink, G. Caldarelli, Phys. Rev. E **76**, 016101 (2007)
[14] P. Grindrod, Phys. Rev. E **66**, 066702 (2002)
[15] This data can be downloaded from the *Eurostat* website: http://epp.eurostat.cec.eu.int
[16] M.-L. Dequeant *et al.*, Science **314**, 1595 (2006)
[17] S. E. Ahnert, K. Willbrand, F. C. S. Brown, T. M. A. Fink, Bioinformatics **22**, 1471 (2006)
[18] P. T. Spellman *et al.*, Molecular Biology of the Cell **9**, 3273 (1998).
[19] List collected from literature, with references, on http://genome-www.stanford.edu/cellcycle/data/rawdata/KnownGenes.doc
[20] I. Simon *et al.*, Cell **106**, 697 (2001).