A clustering coefficient for weighted networks, with application to gene expression data

Article *in* Ai Communications · January 2007

Source: DBLP

CITATIONS
50

READS
203

2 authors, including:

Gabriela Kalna
Beatson Institute for Cancer Research
151 PUBLICATIONS **3,570** CITATIONS

SEE PROFILE

A Clustering Coefficient for Weighted Networks, with Application to Gene Expression Data

Gabriela Kalna [a,*] and Desmond J. Higham [a]

[a] *Department of Mathematics,*
University of Strathclyde,
Glasgow G1 1XH, UK.
E-mail: {ra.gkal,djh}@maths.strath.ac.uk

The clustering coefficient has been used successfully to summarise important features of unweighted, undirected networks across a wide range of applications in complexity science. Recently, a number of authors have extended this concept to the case of networks with non-negatively weighted edges. After reviewing various alternatives, we focus on a definition due to Zhang and Horvath that can be traced back to earlier work of Grindrod. We give a natural and transparent derivation of this clustering coefficient and then analyse its properties. One attraction of this version is that it deals directly with weighted edges and avoids the need to discretise, that is, to round weights up to 1 or down to 0. This has the advantages of (a) retaining all edge weight information, and (b) eliminating the requirement for an arbitrary cutoff level. Further, the extended definition is much less likely to break down due to a 'divide-by-zero'. Using our new derivation and focusing on some special cases allows us to gain insights into the typical behaviour of this measure. We then illustrate the idea by computing the generalised clustering coefficients, along with the corresponding weighted degrees, for pairwise correlation gene expression data arising from microarray experiments. We find that the weighted clustering and degree distributions reveal global topological differences between normal and tumour networks.

Keywords: bioinformatics, computational graph theory, microarray data, network topology, range dependent random graph, small world network, Grindrod-Zhang-Horvath clustering coefficient.

*Corresponding author: Gabriela Kalna, Department of Mathematics, University of Strathclyde, Glasgow G1 1XH, UK.

AI Communications
ISSN 0921-7126, IOS Press.

1. Introduction

Many complex data sets have natural representations as networks. It is accepted that typical real-life networks are neither random graphs in the classical Erdös-Rényi sense nor regular lattices [17,28]. Hence, scientists across a wide range of disciplines face the tasks of summarising, comparing, categorising and modelling these data sets in order to extract meaning and order. Various computational quantities have been used to characterise networks; most prominently the concepts of *pathlength*, *degree* and *clustering coefficient* have proved extremely useful.

Watts and Strogatz [28] coined the phrase *small world network* to describe the commonly occurring situation where a sparse network is highly clustered (like a regular lattice) yet has small pathlengths (like a random graph). Since that landmark paper, many complex networks have been analysed and labelled as small worlds. Similarly, the so-called *scale–free* property of the degree distribution [2,17], has become accepted as a hallmark of many real data sets, although there is now some doubt as to its true prevalence [14,19].

Both the small world and scale–free properties have been widely studied for unweighted, or binary, undirected networks. In the case of more general weighted edges it is of course possible to create a binary network by normalising, imposing a cutoff and rounding to 0 and 1 [21]. However, it is our tenet that the original weights should be respected where possible. While the concept of degree extends readily from unweighted to weighted networks, this is not true for the clustering co-

efficient. A number of authors have recently attempted to generalise the clustering coefficient to the case of weighted edges [3,16,18,29], producing a range of possible definitions.

We present here a natural and transparent derivation of a clustering coefficient for weighted graphs. The resulting definition coincides with those in [8,29] and hence we argue for the use of this Grindrod-Zhang-Horvath clustering coefficient as a generalised measure of clustering. We believe that this measure, along with the corresponding weighted degree distribution, gives an informative high-level picture that can be used for classifying, comparing and modelling weighted networks, just as in the unweighted case. We give some analytical insight into the usefulness of this clustering coefficient by studying the case of range-dependent weights, which is relevant in the biological context. Then we test the definition on an important type of data arising in computational cell biology: gene expression microarray data. Many recent methods for microarray data analysis monitor differences in the expression of genes under various experimental conditions: normal/tumour [5], multiclass cancers [7,20], treatment/survival [23]. Pair-wise gene expression correlation has long been used to predict relationships between genes. Recently, gene co-expression networks have emerged [27,29] connecting genes with high correlation. However, despite the fact that genome-wide gene expression data sets are available, their full potential is often not used and information from only a subset of genes, usually with highest variation, is extracted. Hence, we view these weighted networks as ideal candidates on which to apply the new clustering coefficient framework. Using available microarray data we construct two distinct gene co-expression networks that represent normal and tumour states. We examine weighted clustering coefficients and weighted degree distributions of these networks with the aim of finding tumour-related differences. We emphasize that our aim is to characterize overall network topology rather than to categorize individual genes or samples.

The rest of this article is organised as follows. In section 2 we start with the binary definition of clustering coefficient and list some generalisations that have been proposed for weighted networks. In section 3 we give a natural derivation that leads to the Grindrod-Zhang-Horvath definition, and show how this can be easily computed via matrix products. We then use some simple examples to explore the properties of this coefficient. In section 4 we give some realistic computations on pairwise correlation networks arising from microarray data.

2. Clustering Coefficient and its Generalisations

Consider an undirected graph with normalised weights $0 \leq w_{ij} \leq 1$ between nodes i and j. In the binary case $w_{ij} \in \{0, 1\}$ the clustering coefficient, or curvature, for node k is defined as

$$\text{clust}(k) := \frac{t}{v(v-1)/2}, \tag{1}$$

where v is the number of immediate neighbours of node k, and t is the number of triangles incident to node k [21,28]. In words, clust(k) answers the question "given two nodes that are both connected to node k, what is the likelihood that these two nodes are connected to each other?" It is straightforward to see that the definition breaks down when $v < 2$, that is, node k has less than two immediate neighbours, and otherwise $0 \leq \text{clust}(k) \leq 1$.

Recently, a few different extensions of the clustering coefficient to the general weighted case have emerged. In [16] the weighted clustering coefficient for node k is defined as

$$\text{wclust}_{\text{LF}}(k) := \frac{\sum_{i \neq j \in N(k)} w_{ij}}{v(v-1)},$$

where the term $\sum_{i \neq j \in N(k)} w_{ij}$ can be seen as the total weight of relationship in the neighbourhood $N(k)$ of node k.

Barrat et al.[3] introduced a measure of clustering that combines topological information with the weight distribution of the network

$$\text{wclust}_{\text{B}}\,(\,k) := \frac{1}{s(v-1)} \sum_{i,j} \frac{(w_{ki} + w_{kj})}{2} a_{ik} a_{kj} a_{ij}.$$

Here $s = \sum_j w_{kj}$ denotes the weighted degree of node k and a_{ij} is an element of the underlying binary adjacency matrix. The normalisation factor $s(v-1)$ ensures that $0 \leq \text{wclust}_{\text{B}}(k) \leq 1$. This definition of weighted clustering coefficient considers only weights of edges adjacent to node k but not the weights of edges between neighbours of the node k.

Onnela et al.[18] took into account weights of all edges: adjacent to node k and between-neighbours.

They considered weights $0 \leq w_{ij} \leq 1$ and replaced the number of triangles t in (1) with the sum of triangle intensities

$$\text{wclust}_{\text{O}}(k) := \frac{2\sum_{i,j}(w_{ik}w_{kj}w_{ij})^{1/3}}{v(v-1)}.$$

We remark that the three clustering coefficient definitions above suffer from the drawback that they require an underlying binary network; if this is not available as a separate set of data, then presumably it must be obtained by discretizing the weighted edges. Hence, as in the case where the original binary definition is used for weighted networks [21], they are dependent upon some thresholding parameter. Further, they break down in the case where the number of binary neighbours, ν, is less than 2.

A definition that uses only the network weights was proposed by Zhang and Horvath [29]

$$\text{wclust}_{\text{HZ}}(k) := \frac{\sum_{i\neq k}\sum_{j\neq i,j\neq k} w_{ki}w_{ij}w_{jk}}{(\sum_{i\neq k} w_{ki})^2 - \sum_{i\neq k} w_{ki}^2}. \quad (2)$$

The numerator in (2) was obtained by finding a lower bound for the denominator, this ensuring that $\text{wclust}_{\text{HZ}}$ is in the range $[0, 1]$.

We also mention that rather than one clustering coefficient per node, Schank and Wagner [22] presented a single weighted clustering coefficient for the whole network as

$$\text{wclust}_{\text{S}} := \frac{1}{\sum_v w(v)}\sum_v w(v)c(v).$$

Here $c(v)$ is a clustering coefficient for node v and $w(v)$ a weight function. One possible choice of weight function is the weighted degree.

3. Weighted Clustering

3.1. Definition and Properties

Some simple algebra allows the binary clustering coefficient (1) to be rewritten as

$$\text{clust}(k) := \frac{\sum_{i=1}^{v-1}\sum_{j=i+1}^{v} 1\times 1\times a_{ij}}{\sum_{i=1}^{v-1}\sum_{j=i+1}^{v} 1\times 1}, \quad (3)$$

where $a_{ij} = 1$ if the pairs of neighbours i and j of the node k are connected and $a_{ij} = 0$ otherwise.

Consider now an undirected weighted network of M nodes that is fully connected with weights $0 \leq w_{ij} = w_{ji} \leq 1$ between nodes i and j and $w_{ii} = 0$. Then formula (3) directly extends to the real value case

$$\text{wclust}(k) := \frac{\sum_{i=1}^{M}\sum_{j=1}^{M} w_{ki}w_{kj}w_{ij}}{\sum_{i=1}^{M}\sum_{j=1,j\neq i}^{M} w_{ki}w_{kj}} \quad (4)$$

giving a natural definition for weighted networks. We also mention that the same formula was used in [8] in the context where w_{ij} represents the probability of an edge between nodes i and j in a random network model. Closer inspection shows that the formula (4) has a simple interpretation that is analogous to that of the binary case:

$$\frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{M} w_{ki}w_{kj}w_{ij}$$

is a reasonable measure of the "strength" of triangles that involve node k and

$$\frac{1}{2}\sum_{i=1}^{M}\sum_{j=1j\neq i}^{M} w_{ki}w_{kj}$$

represents the "strength" of the pairs of neighbours that involve node k. Here, "strength" is based on geometric (rather than arithmetic) averaging of the relevant weights. It is easy to verify that (4) retains the property $0 \leq \text{wclust}(k) \leq 1$. It is also trivial to show that (4) approaches the binary value (1) if $w_{ij} \in \{0, 1\}$ are considered.

Computationally, note that the numerator of (4) is

$$\sum_{i=1}^{M} w_{ki}\sum_{j=1}^{M} w_{kj}w_{ij} = \sum_{i=1}^{M} w_{ki}(W^2)_{ki} = (W^3)_{kk}$$

and the denominator is

$$\sum_{i=1}^{M}\sum_{j=1}^{M} w_{ki}w_{kj} - \sum_{i=1}^{M} w_{ki}^2 = (e^T w_k)^2 - ||w_k||_2^2.$$

Here, $(W^p)_{ij}$ denotes the (i, j) element of the pth power of $W \in \mathbb{R}^{M\times M}$, $w_k \in \mathbb{R}^M$ denotes the kth row (or column) of W and $e \in \mathbb{R}^M$ denotes the vector with all elements equal to one. Hence, a neater representation of (4) is

$$\text{wclust}(k) = \frac{(W^3)_{kk}}{(e^T w_k)^2 - ||w_k||_2^2}, \quad (5)$$

which shows that the weighted clustering coefficient can be computed across all nodes in $O(M^3)$ operations. The formula (5) also makes it clear that (4) is entirely equivalent to the Zhang-Horvath definition (2).

Having derived this definition from what we believe to be a natural and informative viewpoint, we now attempt to gain further insights by focussing on particular types of weighted network.

3.2. Limit Forms of Clustering Coefficient

We now zoom to a particular node K of a graph and explore its weighted clustering coefficient (4) in specific cases. Starting with a binary network $w_{ij} \in \{0, 1\}$ we replace zero weights with a small weight $0 < \epsilon \ll 1$ (weak connections) and replace unit weights with $1-\epsilon$ (strong connections). Thus, we relaxed the binary network to a weighted fully connected graph. This is an opposite transformation to a binarisation of a weighted network, when real valued weights below a threshold are mapped to zero and weights above the threshold are mapped to one.

(A) In the first case, let node K have $m > 1$ strong and $n > 1$ weak connections to other nodes in the graph. Then there are (a) $m(m-1)/2$ strong-strong, (b) mn strong-weak and (c) $n(n-1)/2$ weak-weak neighbour pairs. Let there be r, s and u strong edges between neighbours in cases (a), (b) and (c) respectively. It is easy to show that equation (4), for $\epsilon \to 0$, results in $\text{wclust}(K) = 2r/m(m-1)$. In words, r strong triangles are built over $m(m-1)/2$ strong neighbour pairs. Thus the weighted clustering coefficient (4) approaches the binary value (1).

(B) In the second case we consider the marginal setting $v = 1$: node K has a strong connection, $1-\epsilon$, only to one node P and n weak, ϵ, connections to all other nodes in a complete graph. Then n out of all possible neighbour pairs involve the strong edge between nodes K and P and $n(n-1)/2$ pairs are formed by n weak edges adjusted to node K. Between-neighbour edges will be again strong or weak. Let there be r strong edges with one end in node P and s strong edges between "weakly" connected neighbour nodes of node K. Then from (4) we get $\text{wclust}(K) = (r\epsilon(1-\epsilon)^2 + (n-r)\epsilon^2(1-\epsilon) + s\epsilon^2(1-\epsilon) + (n(n-1)/2 - s)\epsilon^3)/(n\epsilon(1-\epsilon) + n(n-1)\epsilon^2/2)$. This expression results in r/n for $\epsilon \to 0$. In words, we can get to r out of n "weakly" connected neighbours of the node K through the strong edge KP and strong edges connecting node P with these r nodes. It is clear that $\text{wclust}(k) = 1$ only if $r = n$, that means there is a strong edge between P and all nodes weakly connected to K. Because $\text{wclust}(k) = 0$ if $r = 0$, the strong edge between nodes K and P is the only edge involving node P. That means this edge would be separated from the graph in the corresponding discretised network.

Case **B** reveals an important advantage of the generalised definition (4). It continues to provide useful information in the small ϵ regime where *any discretization process based on thresholding to a binary network would result in $v = 1$ and hence an undefined clustering coefficient in (1).*

3.3. Uniform Weights

Another case where the clustering coefficient simplifies arises when node K has equal weights with all other nodes: $w_{Kj} = w = \text{constant}$ for all $j \neq K$. In this case we have

$$\text{wclust}(K) = \frac{w^2 \sum_{i=1}^{M} \sum_{j=1}^{M} w_{ij}}{w^2 \sum_{i=1}^{M} \sum_{j=1, j\neq i}^{M} 1} = \frac{\sum_{i=1}^{M-2} \sum_{j=1}^{M-1} w_{ij}}{(M-2)(M-1)}$$

and we see that $\text{wclust}(K)$ then reflects the average connectivity between the other nodes in the network. In particular, if all between-neighbour connections are equal, i.e. $w_{ij} = w = \text{constant}$ for all $i \neq j \neq K$, (4) results in

$$\text{wclust}(K) = \frac{w \sum_{i=1}^{M} \sum_{j=1, j\neq i}^{M} w_{ki} w_{kj}}{\sum_{i=1}^{M} \sum_{j=1, j\neq i}^{M} w_{ki} w_{kj}} = w.$$

Generally, the weighted clustering coefficient can be bounded in terms of the extremal network coefficients,

$$\min_{i\neq j\neq K}(w_{ij}) \leq \text{wclust}(K) \leq \max_{i\neq j\neq K}(w_{ij}).$$

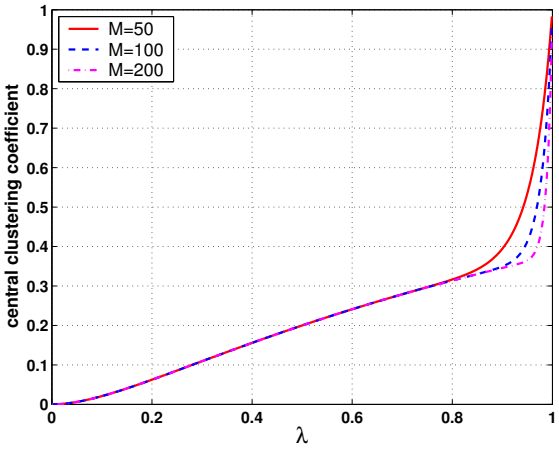

Fig. 1. Clustering coefficient (4) of the central node in the weighted graph defined by (6).

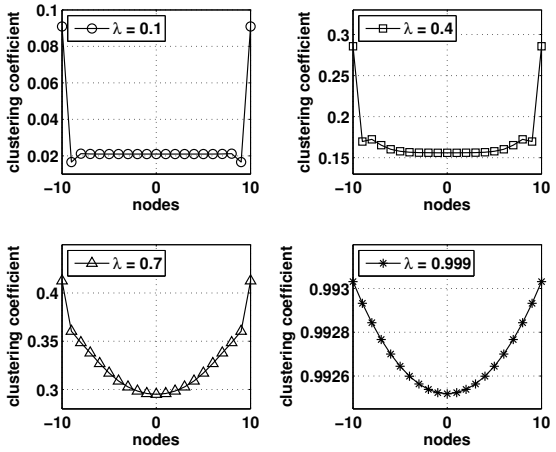

Fig. 2. Clustering coefficient of 21 nodes in the weighted graph defined by (6) for different values of λ.

3.4. Range Dependent Weights

The concept of a *range-dependent weighted random graph*, or RENGA, was introduced and analyzed by Grindrod [8]. These graphs were further studied by [10] and were used in [13] to produce test matrices for linear algebra software. The idea of using a distance metric to induce edges has proved successful in the modelling of real networks, especially in biology [19]. We may adapt the RENGA idea to the case of non-random range-dependent weights. Suppose that the nodes are ordered $-M, \ldots, -1, 0, 1, \ldots, M$ and that the connectivity weight decays as a function of lattice distance. To be specific, we let

$$w_{ij} = w_{ji} = \lambda^{|i-j|}, \tag{6}$$

for some $\lambda \in [0, 1]$. At one extreme, $\lambda \approx 0$, there are no edges after discretising to a binary network, and hence the traditional clustering coefficient is undefined. At the other extreme, $\lambda = 1$, all edges are present after discretising to a binary network, and hence the traditional clustering coefficient is 1 for each node. In Figure 1 we use networks of size $M = 50, 100, 200$ and compute the generalised clustering coefficient (4) for the central node, $k = 0$, as λ ranges from 0 to 1. Note that the definition (4) makes sense for any $\lambda > 0$. We see that the clustering coefficient approaches the value zero as λ approaches zero from above; this is perfectly reasonable behaviour. Further, as λ increases away from zero, the clustering coefficient monotonically increases, and it matches the binary value of 1 at $\lambda = 1$.

Figure 2 shows the clustering coefficient of all nodes in the case $M = 10$ for $\lambda = 0.1, 0.4, 0.7$ and 0.999. For $\lambda = 0.1$ and $\lambda = 0.4$ the clustering coefficient is not monotonic in the nodal distance from the boundary, and there is a dramatic difference between the neighbouring nodes at positions $M - 1$ and M. The pictures suggest that there is no $M \to \infty$ "continuum limit" in the sense that the clustering coefficients do not appear to lie on a smooth curve. Because of the special structure of the weights, we are able to investigate this issue analytically.

Some straightforward analysis produces the formulas

$$\begin{aligned} \text{wclust}(M) &= \frac{\lambda \sum_k k\lambda^{2(k-1)}}{(1+\lambda)\sum_k k\lambda^{2(k-1)}} \\ &\approx \frac{\lambda}{(1+\lambda)}, \end{aligned}$$

$$\text{wclust}\,(M-1) = \frac{\sum_{k=1}(k+1)\lambda^{2k}}{1+(1+\lambda)\sum_k (k+1)\lambda^{2k-1}},$$

$$\text{wclust}\,(M-2) = \frac{\sum_k (k+2)\lambda^{2k}}{1+(1+\lambda)\big(1+\sum_k (k+2)\lambda^{2k-1}\big)},$$

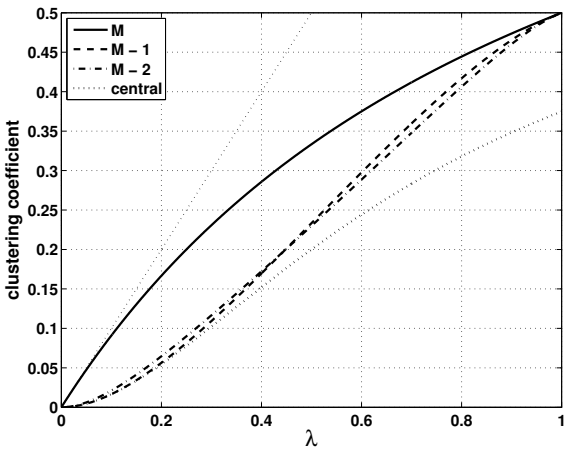

Fig. 3. Clustering coefficient of the three marginal nodes and the central node 0 in the infinite ($M \to \infty$) weighted graph defined by (6).

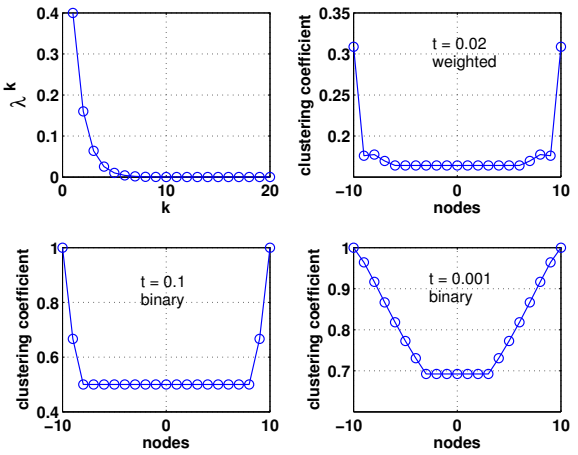

Fig. 4. Clustering coefficient: thresholding to weighted and binary networks.

$$\text{wclust}\,(0) = \frac{3\lambda^2 \sum_k k\lambda^{2(k-1)}}{1+\sum_{k=1}\lambda^{2k-1}\big(4k+(4k+1)\lambda\big)}.$$

By summing appropriate geometric series we get, for fixed $0 \le \lambda < 1$, the following formulas and corresponding ranges of values for the weighted clustering coefficients in the limit $M \to \infty$

$$\text{wclust}(M) = \frac{\lambda}{1+\lambda}, \qquad \text{wclust}(M) \in [0, 1/2],$$

$$\text{wclust}(M-1) = \frac{\lambda^2(2-\lambda^2)}{(1+\lambda)(1+\lambda-\lambda^2)}, \qquad \text{wclust}(M-1) \in [0, 1/2],$$

$$\text{wclust}\,(M-2) = \frac{\lambda^2(3-2\lambda^2)}{(1+\lambda)(1+3\lambda-2\lambda^2-2\lambda^3+\lambda^4)}, \qquad \text{wclust}(M-2) \in [0, 1/2],$$

$$\text{wclust}(0) = \frac{\lambda^2(2+\lambda^2)}{(1+\lambda^2)(1+\lambda)^2}, \qquad \text{wclust}(0) \in [0, 3/8].$$

The limit for the endpoint node, wclust(M), is exceptional in that, as a function of λ, the expression has a tangent of slope $= 1$ at the origin. The other cases have tangents of slope $= 0$. We also note that letting $\lambda \to 1$ in these expressions does not reproduce the clustering coefficient value that arises from inserting $\lambda = 1$ into (6). In Figure 3 we illustrate these effects.

Binarising these networks produces very different behaviour. Figure 4 looks at the the case $M = 10$ and $\lambda = 0.4$ (cf. the upper right picture in Figure 2). The range of weights appearing in the network, $\{\lambda^k\}$, are shown for reference in the upper left picture. The lower left and right pictures show the clustering coefficients when weights below t are mapped to zero and weights above t are mapped to one, for $t = 0.1$ and $t = 0.001$. The upper right picture, which more closely resembles the fully weighted version, corresponds to zeroing weights below t and leaving weights above t unchanged.

Overall, in analysing the interesting, nontrivial, range-dependent case (6), we see again that respecting the real valued weights can produce a significantly different and more meaningful picture, compared with discretising.

4. Microarray Illustration

We now examine the distribution of the clustering coefficient (4) in practice, along with that of the corresponding weighted degree, using pairwise

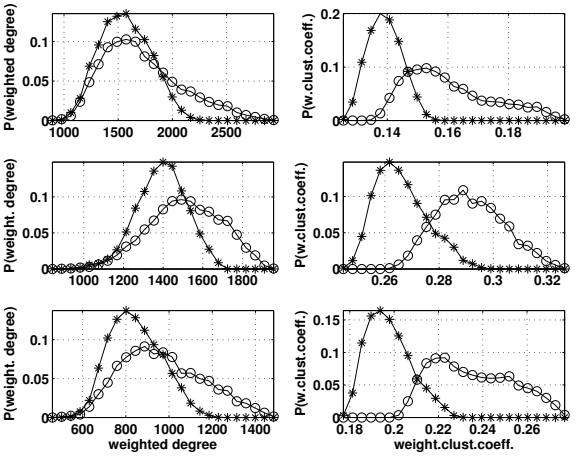

Fig. 5. Probability of weighted degree (left) and clustering coefficient (right). Liver cancer (top), breast cancer (middle) and lymphoma (bottom): normal (circles) and tumour (stars).

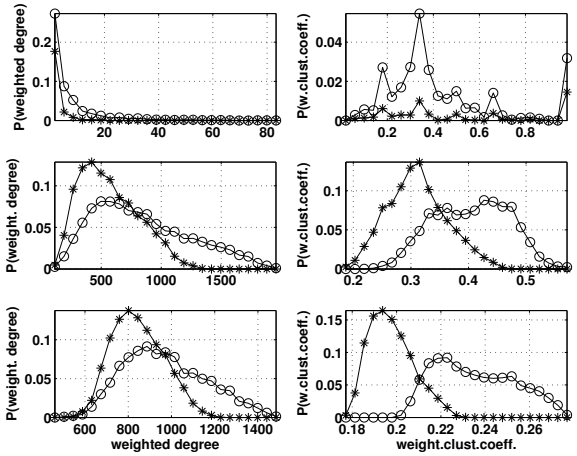

Fig. 6. Probability of weighted degree (left) and clustering coefficient (right). Lymphoma: normal (circles) and tumour (stars). Binary network from threshold = 0.8 (top), threshold P value 0.05 (middle) and weighted network (bottom).

correlation networks arising from cDNA microarray data. Most importantly, we would like to explore differences in character of weighted degree and clustering coefficient distributions of two different networks: normal and tumour.

cDNA microarrays are a genome-scale technology used for assessing differential gene expression [4,24]. Two different samples labeled with different fluorescent dyes hybridize on the same array. One sample is prepared from a reference mRNA and the other from mRNA isolated from the experimental cells. The common reference, or universal control, is collected from a pool of cell lines or a mix of all analysed samples [26,9]. The initial data arising from cDNA microarray experiments are relative mRNA levels – experiment to control ratios. The data from different arrays require complex normalisation prior to comparison. Normalised values are usually organised in the form of a rectangular $M \times N$ matrix of log-transformed ratios a_{ij} of $i = 1, \ldots, M$ genes in a set of $j = 1, \ldots, N$ samples.

To build weighted networks for further analysis of microarray data a suitable similarity measure needs to be chosen which produces real value results and has a reasonable biological interpretation. We consider the Pearson correlation

$$\mathrm{cor}(i,j) = \frac{\sum_{k=1}^{N}(a_{ik}-\mu_i)(a_{jk}-\mu_j)}{\sigma_i \sigma_j},$$

where μ_i and σ_i are respectively the mean and the standard deviation of gene i log-ratios, as a measure of similarity between the gene expression profiles. This measure, or similar correlation measures, proved useful for analysing microarray data in [12,15,21,23,29]. We define pairwise gene similarity weights $w_{ij} = |\mathrm{cor}(i,j)|$, for $1 \le i,j \le M$, with $w_{ij} = w_{ji} \in [0,1]$ and $w_{ii} = 0$. A large weight w_{ij} indicates that genes i and j are highly co-expressed (or anti-expressed). In this representation the $M \times M$ matrix W denotes the symmetric weight matrix encoding the strength of connection between pairs of genes.

Aware of the fact that different numbers of genes as well as samples in data sets can affect values of correlation and consequently distort comparisons of both weighted degrees and clustering coefficients, we looked for data consisting of the same number of normal and tumour samples for the same set of genes. In this experiment we used cDNA microarray data for normal and tumour tissues from [6]. Data processing performed by those authors included filtering of genes with more than 70% missing values or less than 4 observations, UniGene mapping, and imputation of missing values. The original data can be downloaded from the Stanford Microarray Database.

We selected data sets with more than ten samples in normal and tumour subsets. We present results for liver cancer (12065 genes, 76 samples), breast cancer (5603 genes, 13 samples), and lymphoma (4615 genes, 31 samples) [5,25,1]. Figure 5 shows the distribution of the weighted cluster-

ing coefficient (right), and also the distribution of the weighted degree (left) arising from these data. Circle-line and star-line represent the distributions of normal and tumour networks respectively.

We emphasize that our aim is to study the 'big-picture' issue of overall network topology, as opposed to the 'fine-detail' issue of clustering individual genes and/or samples [11,15]. Figure 5 reveals global topological differences between the two network types. We have performed the Kolmogorov Smirnov test (kstest2 in MATLAB) on the weighted degree and weighted clustering coefficient values of normal and tumour networks. The test confirmed that normal and tumour distributions are highly significantly different ($P < 0.001$). In general the tumour samples give rise to smaller and more peakily distributed clustering and degree. Degree ranges of normals and tumours start from a similar value but the degree range of tumours is narrower. Large numbers of genes in normal samples show a high degree of connection to other genes. Differences in clustering coefficient distributions are more striking. Distribution ranges of normal and tumour networks only partly overlap: most genes in normal networks have higher clustering coefficient than any gene in tumour networks. One possible biological interpretation of these results is that the diseased state is consistent with a breakdown of the control mechanisms that regulate the co-expression of functionally related genes. Similar results, supported by Kolmogorov Smirnov test ($P < 0.001$), arose when the matrix $|A|$ was used instead of A. In this case correlation between genes is based on overall activity, with over and under expression both regarded as equally valid evidence of a gene leaving its typical state.

Given that the weighted clustering coefficient produces interesting results, it is pertinent to ask whether careful thresholding to a discretised binary network [21] can also reproduce these findings. Clearly there is a whole parameterized family of such binary networks. In particular, high thresholding may exclude interesting features of the networks. For example, when weights above the threshold of 0.8 are re-set to 1 and the remaining weights are re-set to zero, the clustering coefficient and weighted degree distributions could not reveal the differences observed from original networks; see Figure 6 top.

For a more systematic approach, P values may be used to decide on significance of correlation. Even in this case, however, somewhat arbitrary thresholds must be imposed. For the lymphoma networks, suppose we take the view that correlations ≥ 0.355 are significant (corresponding to $P \leq 0.05$) and correlations ≥ 0.456 are highly significant (corresponding to $P \leq 0.01$). This would mean that only 18% (12%) of all possible edges are significant and 8% ($< 5\%$) are highly significant in the normal (tumour) lymphoma network, so that a large amount of data is being discarded. (Of course, there are computational benefits from introducing sparsity, but for the network sizes in these experiments this is not a significant issue.) In the middle of Figure 6 we plot data for the $P \leq 0.05$ binary networks. Comparing this result with the resalt from the original weighted network (bottom part of Figure 6), we see that very similar topology is revealed. We conclude that the weighted clustering coefficient approach, which does not require the use of arbitrary parameters and does not discard data, automatically produces results consistent with those arrived at through experimenting with the parameter-dependent P value version.

5. Summary

Our aim here was to argue that out of the possible ways that have been proposed to generalise the clustering coefficient to the case of a weighted network, there is one very promising candidate; namely the Grindrod-Zhang-Horvath version [8,29]. We gave a natural derivation and illustrated its behaviour on specific classes of network. Particular advantages of the definition are:

- It is a true generalisation, collapsing smoothly to the binary case when edge weights tend to $\{0, 1\}$ values.
- It can provide meaningful results in cases where any type of binary thresholding produces break-down.
- It reveals natural topological properties of real networks, and can do this without the need to specify parameters or discard potentially useful data.
- For microarray data, the weighted clustering coefficient coefficient lead to a clear hypothesis about the difference between normal and cancerous states.

Acknowledgements

Both authors are supported by EPSRC grants GR/S62383/01 and EP/EO49370/1.

References

[1] A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. H. Jr, L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Botstein, P. O. Brown, and L. M. Staudt, Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling, *Nature* **403** (2000), 503–511.

[2] A. Barabasi and R. Albert, Emergence of scaling in random networks, *Science* **286** (1999), 509–512.

[3] A. Barrat, M. Barthélemy, R. Pastor-Satorras, and A. Vespignani, The architecture of complex weighted networks, *PNAS* **101** (2004), 3747–3752.

[4] P. Brown and D. Botstein, Exploring the new world of the genome with dna microarrays, *Nature Genet.* **21** (1999), 33–37.

[5] X. Chen, S. Cheung, S. So, S. Fan, C. Barry, J. Higgins, K. Lai, J. Ji, S. Dudoit, I. Ng, M. Van De Rijn, and P. Botstein, D. Brown, Gene expression patterns in human liver cancers, *Molecular Biology of the Cell* **13** (2002), 1929–1939.

[6] J. Choi, U. Yu, O. Yoo, and S. Kim, Differential coexpression analysis using microarray data and its application to human cancer, *Bioinformatics* **21** (2005), 4348–4355.

[7] T. Golub, D. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. Mesirov, H. Coller, M. Loh, J. Downing, M. Caliguri, C. Bloomfield, and E. Lander, Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science* **286** (1999), 531–537.

[8] P. Grindrod, Range-dependent random graphs and their application to modeling large small-world proteome datasets, *Physical Review E* **66** (2002), 066702.

[9] G. Hardiman, Microarray platforms - comparisons and contrasts, *Pharmacogenomics* **5** (2004), 487–502.

[10] D. J. Higham, Spectral reordering of a range-dependent weighted random graph, *IMA Journal of Numerical Analysis* **25** (2005), 443–457.

[11] D. J. Higham, G. Kalna, and M. Kibble, Spectral clustering and its use in bioinformatics, *J.Computat. Appl. Math.* **204** (2007), 25–37.

[12] D. J. Higham, G. Kalna, and K. Vass, Spectral analysis of two-signed microarray expression data, *Mathematical Medicine and Biology* **24** (2007), 131–148.

[13] Y. Hu and J. A. Scott, Hsl_mc73: A fast multilevel fiedler and profile reduction code, Technical Report RAL-TR-2003-036, Rutherford Appleton Laboratory, Oxfordshire, 2003.

[14] R. Khanin and E. Wit, How scale-free are gene networks?, *J. Comput. Biol* **13** (2006), 810–818.

[15] Y. Kluger, R. Basri, J. Chang, and M. Gerstein, Spectral biclustering of microarray data: coclustering genes and conditions, *Genome Research* **13** (2003), 703–716.

[16] L. Lopez-Fernandez, G. Robles, and J. Gonzalez-Barahona, Applying social network analysis to the information in cvs repositories, in: *Proc. of the 1st Intl. Workshop on Mining Software Repositories (MSR2004)*, (2004), pp. 101–105.

[17] M. Newman, The structure and function of complex networks, *SIAM Review* **45** (2003), 167–256.

[18] J.-P. Onnela, J. Sarami, J. Kertz, and K. Kaski, Intensity and coherence of motifs in weighted complex networks, *Phys. Rev. E* **71** (2005), 065103.

[19] N. Pržulj, D. Corneil, and I. Jurisica, Modeling interactome: Scale-free or geometric?, *Bioinformatics* **20** (2004), 3508–3515.

[20] S. Ramaswamy, P. Tamayo, R. Rifkin, S. Mukherjee, C.-H. Yeang, M. Angelo, C. Ladd, M. Reich, E. Latulippe, J. Mesirov, T. Poggio, W. Gerald, M. Loda, E. Lander, and T. Golub, Multiclass cancer diagnosis using tumor gene expression signatures, *PNAS* **98** (2001), 15149–15154.

[21] J. Rougemont and P. Hingamp, Dna microarray data and contextual analysis of correlation graphs, *BMC Bioinformatics* **4** (2003), 4–15.

[22] T. Schank and D. Wagner, Approximating clustering coefficient and transitivity, *J. of Graph Algorithms and Applications* **9** (2005), 265–275.

[23] M. Segal, Microarray gene expression data with linked survival phenotypes: Diffuse large-b-cell lymphoma revisited, *Biostatistics* **7** (2006), 268–285.

[24] M. Shena, D. Shalon, R. David, and P. Brown, Quantitative monitoring of. gene expression patterns with a complementary dna microarray, *Science* **270** (1995), 467–470.

[25] T. Sorlie, C. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M. Eisen, M. van de Rijn, S. Jeffrey, T. Thorsen, H. Quist, J. Matese, P. Brown, D. Botstein, P. E. Lonning, and A. Borresen-Dale, Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications, *PNAS* **98** (2001), 10869–10874.

[26] E. Sterrenburg, R. Turk, J. Boer, G. van Ommen, and J. den Dunnen, A common reference for cdna microarray hybridizations, *Nucleic Acids Res.* **30** (2002), e116.

[27] J. Stuart, E. Segal, D. Koller, and S. Kim, A gene-coexpression network for global discovery of conserved genetic modules, *Science* **302** (2003), 249–255.

[28] D. Watts and S. Strogatz, Collective dynamics of small-world networks, *Nature* **393** (1998), 440–442.

[29] B. Zhang and S. Horvath, A general framework for weighted gene co-expression network analysis, *Statistical Applications in Genetics and Molecular Biology* **4**, (2005).