PHYSICAL REVIEW E **71**, 065103(R) (2005)

Intensity and coherence of motifs in weighted complex networks

Jukka-Pekka Onnela,[1] Jari Saramäki,[1] János Kertész,[1,2] and Kimmo Kaski[1]
[1]*Laboratory of Computational Engineering, Helsinki University of Technology, Espoo, Finland*
[2]*Department of Theoretical Physics, Budapest University of Technology and Economics, Budapest, Hungary*
(Received 13 September 2004; revised manuscript received 23 December 2004; published 13 June 2005)

The local structure of unweighted networks can be characterized by the number of times a subgraph appears in the network. The clustering coefficient, reflecting the local configuration of triangles, can be seen as a special case of this approach. In this paper we generalize this method for weighted networks. We introduce subgraph "intensity" as the geometric mean of its link weights and "coherence" as the ratio of the geometric to the corresponding arithmetic mean. Using these measures, motif scores and clustering coefficient can be generalized to weighted networks. To demonstrate these concepts, we apply them to financial and metabolic networks and find that inclusion of weights may considerably modify the conclusions obtained from the study of unweighted characteristics.

DOI: 10.1103/PhysRevE.71.065103 PACS number(s): 89.75.Hc, 87.16.Ac, 89.65.−s

The network approach to complex systems has turned out to be extremely fruitful and it has revealed some general principles applicable to a large number of systems. Studies have produced unexpected findings such as the ubiquity of scale freeness, the frequent appearance of high clustering, and the relationship between functionality and the high appearance frequency of specific motifs. This approach has also led to a number of novel paradigmatic models, providing a holistic framework in which the details of the interactions between the constituents of the complex systems are disregarded and only their scaffolds are considered [1].

A deeper understanding of these systems requires that, in addition to the underlying network structure, information about the strength of interactions is also taken into account. This is accomplished by assigning weights to the links, such as transportation fluxes in the Internet and air traffic networks [2,3], or the reaction fluxes building the metabolic pathways of a cell [4]. Weights can also be obtained by applying a classification (or clustering) scheme to a correlation matrix, or for understanding the structure underlying the dynamics of microarray [5] and stock market data [6,7]. Optimal paths [8] and minimum spanning trees [9] also clearly depend on the distribution of weights. These examples indicate the need to generalize the network characteristics to weighted networks. Some recent efforts towards this goal are the discussion of the clustering coefficient for node weights [10], introduction of a definition for the link weighted case [3,11], and the mapping of weighted networks to multigraphs [12]. Our aim in this paper is to introduce a set of practical tools that may be used to study the structure of a diverse group of systems where interactions strengths can be obtained and where omitting them would lead to a considerable loss of information. Many biological and social systems are expected to fall into this category.

In general, we consider any weighted network as a fully connected graph where some of the links bear zero weights. For simplicity, we deal with (directed or undirected) networks where the weight w_{ij} between nodes i and j is nonnegative and not necessarily normalized. We introduce the intensity $I(g)$ of subgraph g with vertices v_g and links l_g as the geometric mean of its weights

$$I(g) = \left(\prod_{(ij) \in l_g} w_{ij} \right)^{1/|l_g|}, \tag{1}$$

where $|l_g|$ is the number of links in l_g. The definition suggest a shift in perspective from regarding subgraphs as discrete objects (either exist or not) to a continuum of subgraph intensities, where zero or very low intensity values imply that the subgraph in question does not exist or exists at a practically insignificant intensity level. In practice, low intensity values could result, for example, from measurement noise.

Due to the nature of the geometric mean, the subgraph intensity $I(g)$ may be low because one of the weights is very low, or it may result from all of the weights being low. In order to distinguish between these two extremes, we introduce subgraph coherence $Q(g)$ as the ratio of the geometric to the arithmetic mean of the weights as

$$Q(g) = I(g)|l_g| \Big/ \sum_{(ij) \in l_g} w_{ij}. \tag{2}$$

Here $Q \in [0,1]$ and it is close to unity only if the subgraph weights do not differ much, i.e., are internally coherent.

The concept of a motif was originally introduced to denote "patterns of interconnections occurring in complex networks at numbers that are significantly higher than those in randomized networks" [13]. However, this has led to some confusion, which partly stems from the specification of the random ensemble, i.e., the underlying null hypothesis [14]. We define a motif as a set (ensemble) of topologically equivalent subgraphs of a network. With weighted networks it becomes more natural to deal with intensities as opposed to numbers of occurrence, where the latter is obtained as a special case of the former. The motifs showing statistically significant deviation from some reference system can then be called high or low intensity motifs.

We define the total intensity I_M of a motif M in the network as the sum of its subgraph intensities $I_M = \Sigma_{g \in M} I(g)$. For certain weighted directed motifs, the total intensities can be computed using simple matrix operations. Let the $N \times N$ weight matrix **W** describe the network weights. Analogously, let **A** represent the underlying $N \times N$ adjacency matrix such

that $a_{ij}=1$ if $w_{ij}>0$, and $a_{ij}=0$ if $w_{ij}=0$. In an unweighted network, the number of directed paths returning to the starting node after k steps can be written as

$$N(k)=\sum_{i_1,\ldots,i_k}\prod_{x=1}^{k}a_{i_x,i_{x+1}}=\mathrm{Tr}\{\mathbf{A}^k\}, \tag{3}$$

where the summation goes over all possible sites and $i_{k+1}=i_1$ [1]. Let $\mathbf{W}^{(1/k)}$ represent a matrix obtained from $\mathbf{W}=[w_{ij}]$ by taking the kth root of its individual elements such that $\mathbf{W}^{(1/k)}=[w_{ij}^{1/k}]$. The total intensity of motif M in the network is

$$I_M=a_M\sum_{i_1,\ldots,i_k}\left(\prod_{x=1}^{k}w_{i_x,i_{x+1}}\right)^{1/k}=a_M\mathrm{Tr}\{(\mathbf{W}^{(1/k)})^k\}, \tag{4}$$

where a_M is a combinatorial factor ensuring that each subgraph is counted only once. For example, for the nonfrustrated triangle (Fig. 3, middle column) the total intensity becomes $I_\Delta=\frac{1}{3}\mathrm{Tr}\{(\mathbf{W}^{(1/3)})^3\}$. A change in the direction of a link can be taken into account using the matrix transpose. For some motifs, such as the path of order 2 (Fig. 3, left column) we need a "block" matrix $\mathbf{B}=[b_{ij}]$ to prevent us from double-counting subgraphs, in this case to prevent counting every triangle also as three paths of order 2. In this matrix the diagonal elements $b_{ii}=0$ and for the nondiagonal elements $b_{ij}=0$ when $a_{ij}=1$ or $a_{ji}=1$, and otherwise $b_{ij}=1$. This allows us to write the total intensity of the path motif as $I_\angle=\mathrm{Tr}\{\mathbf{W}^{(1/2)}\mathbf{W}^{(1/2)}\mathbf{B}\}$. We prevent double counting here for reasons of compatibility with earlier work, but find that it poses no serious problem as long as the system of counting is systematically applied both in the empirical and random case. Double counting could, in fact, be desirable if the interaction strength measurements are noisy. For example, envision adding a small number ϵ to every link (including the zeros) to represent a noise component. This would lead to large number of "false positive" triangles in the network, and counting only them would lead us to miss the structure hidden inside them, e.g. important paths of order 2.

In Ref. [13] the z score for studying the statistical significance of motif occurrences was defined as

$$z_M=(N_M-\langle n_M\rangle)/\sigma_M, \tag{5}$$

where N_M is the number of subgraphs in motif M in the empirical network and $\langle n_M\rangle$ is the expectation of their number in the reference ensemble, and σ_M is the standard deviation of the latter. Replacing the number of subgraphs by their intensities generalizes the z score to motif intensity score

$$\tilde{z}_M=(I_M-\langle i_M\rangle)/(\langle i_M^2\rangle-\langle i_M\rangle^2)^{1/2}, \tag{6}$$

where i_M is the total intensity of motif M in one realization of the reference system. It is clear that Eqs. (5) and (6) coincide for binary weights, implying that $\tilde{z}\to z$ in the limit. As an analog to the motif intensity score, we introduce the motif coherence score as

$$\tilde{z}'_M=(Q_M-\langle q_M\rangle)/(\langle q_M^2\rangle-\langle q_M\rangle^2)^{1/2}, \tag{7}$$

where Q_M and q_M are the total coherence for motif M in the empirical network and in one realization of the reference system, respectively. As the coherence of an unweighted subgraph is unity, also $\tilde{z}'\to z$ as the weights become binary.

Triangles are among the simplest nontrivial motifs and they play an important role as one of the basic quantities of network characterization in defining the clustering coefficient C_i at node i as

$$C_i=\frac{2t_i}{k_i(k_i-1)}, \tag{8}$$

where k_i is the degree of node i and t_i is the number of triangles attached to the node [1,15]. This quantity is normalized between 0 and 1, and it characterizes the tendency of the nearest neighbors of node i to be interconnected.

As triangles are one type of subgraph, the definition in Eq. (1) may be used to yield the weighted clustering coefficient $\tilde{C}_i$ by replacing the number of triangles t_i in Eq. (8) with the sum of triangle intensities as

$$\tilde{C}_i=\frac{2}{k_i(k_i-1)}\sum_{j,k}(\tilde{w}_{ij}\tilde{w}_{jk}\tilde{w}_{ki})^{1/3}, \tag{9}$$

where we use weights scaled by the largest weight in the network, $\tilde{w}_{ij}=w_{ij}/\max(w_{ij})$. This definition fulfills the requirement that $\tilde{C}_i\to C_i$ as the weights become binary. We can relate the unweighted and weighted clustering coefficients through the average intensity of triangles at node i as $\overline{I_i}=(1/t_i)\Sigma_{g\in\mathcal{N}(v_i)}I(g)$, where $\mathcal{N}(v_i)$ denotes the neighborhood of node i, and this allows us to write the weighted clustering coefficient as

$$\tilde{C}_i=\overline{I_i}C_i. \tag{10}$$

This equation gives a plausible interpretation of the weighted clustering coefficient: It is the unweighted (topological) clustering coefficient renormalized by the average intensity of triangles at the node. Naturally, a weighted clustering coefficient $\tilde{C}'_i$ can also be formulated by renormalizing the unweighted coefficient by the average coherence $\overline{Q_i}$, instead of the average intensity $\overline{I_i}$, around node i.

An alternative definition for weighted clustering coefficient was given in Ref. [3] as

$$\hat{C}_i=\frac{1}{s_i(k_i-1)}\sum_{j,k}\frac{(w_{ij}+w_{ik})}{2}a_{ij}a_{ik}a_{jk}, \tag{11}$$

where s_i denotes the strength of node i, defined as $s_i=\Sigma_j w_{ij}$, and a_{ij} is an element of the underlying binary adjacency matrix. This definition considers only two of the three link weights, namely, those adjacent to node i (w_{ij} and w_{ik}) and requires that a link exist also between nodes j and k but does not take its weight (w_{jk}) into account. The difference between the two weighted clustering coefficients $\tilde{C}_i$ and $\hat{C}_i$ is illustrated schematically in Fig. 1.

Next we apply these concepts to two real networks.

(A) *Undirected financial network*. We considered a set of daily price data for $N=477$ NYSE traded stocks from 1980 to 2000. We calculated the correlation matrix by extracting four-year return windows in order to study the system's dy-

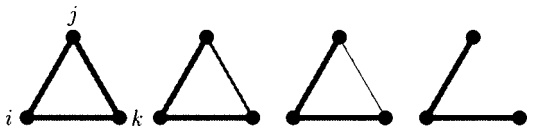

FIG. 1. A schematic illustration of the difference between $\hat{C}_i$ and $\tilde{C}_i$. The weight w_{jk} is gradually decreased from left to right. The value of $\hat{C}_i$ is equal for the first three triangles and drops to zero suddenly for the fourth triangle as $w_{jk} \to 0$, implying that $a_{jk}=0$. In contrast, the value of $\tilde{C}_i$ decreases as $C_i \sim w_{jk}^{1/3}$, tending smoothly to zero in the limit.

namics. Here the nodes correspond to stocks, and the weighted undirected links to the elements in the correlation matrix. Thus, the stronger the weight, the stronger the coupling between the stock returns in terms of their linear correlation. The links are inserted in the network in descending order starting from the strongest one until a predetermined number of links has been reached. The method is described in detail in Ref. [7].

We have shown earlier that the famous Black Monday (10/19/1987) causes a temporary transition not only in the topology but also in the weights of the network [16]. Our aim is to use it as an example of a network undergoing this type of twofold transition (topology and weights) and to see whether the changes are reflected in the network's clustering statistics. In Fig. 2 we show the three clustering coefficients, averaged over the network, as functions of time: the unweighted C of Eq. (8), the weighted $\hat{C}$ introduced in Ref. [3] and given in Eq. (11), and the weighted $\tilde{C}$ introduced in Eq. (9).

The crash is not seen very clearly in C, as it can only capture the topological aspects of the transition. The weighted coefficient $\hat{C}$ is also fairly insensitive to the changes in link weights and practically coincides with C. The fact that $\tilde{C}$ does reflect the transition indicates its ability to capture both aspects of the transition. The average values

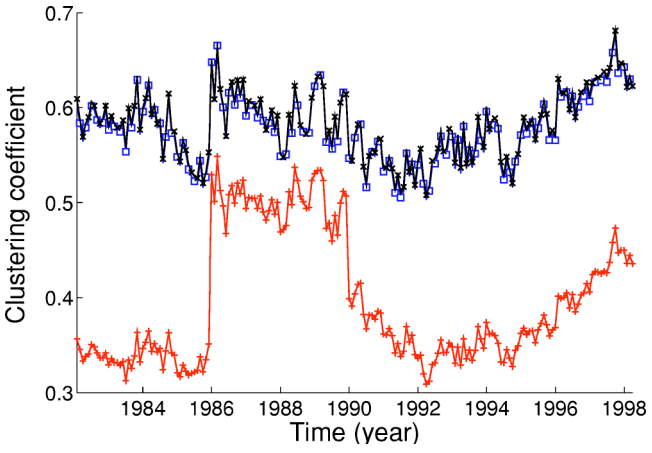

FIG. 2. (Color online) Average clustering coefficients for the financial network. The weighted clustering coefficient $\tilde{C}$ (+) of Eq. (9) shows the effect of Black Monday clearly. The unweighted C (□) of Eq. (8) and the weighted $\hat{C}$ (×) of Eq. (11) practically coincide (the markers □ and × are used alternately).

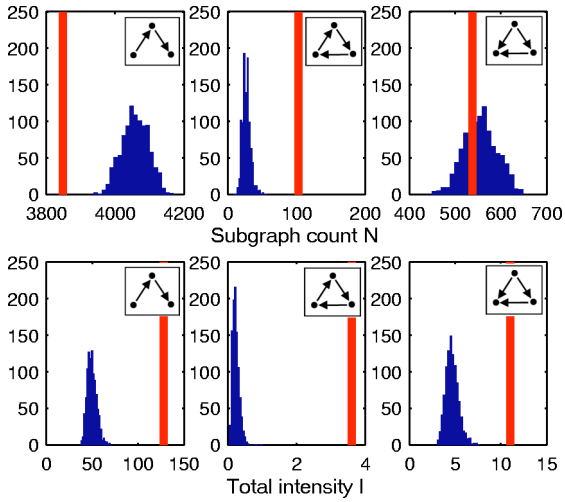

FIG. 3. (Color online) Motif intensities for the empirical network (vertical lines) and the corresponding random ensembles (histograms), for the unweighted (upper panel) and weighted (lower panel) cases.

for the clustering coefficients outside (inside) the crash "period" are $C=0.57$ ($C=0.60$), $\hat{C}=0.58$ ($\hat{C}=0.60$), and $\tilde{C}=0.36$ ($\tilde{C}=0.50$). These numbers imply that C and $\hat{C}$ increase less than 5% during the crash which is less than their normal (outside the crash period) fluctuation, measured at 6.2% as their standard deviation relative to the mean. However, the crash increases $\tilde{C}$ by 39%, which is considerably larger than the the level of fluctuation at 9.7%. Thus, $\tilde{C}$ has a considerably higher "signal-to-noise" ratio. The results are not affected significantly by the value of the predetermined threshold. In the limit of inserting all the links of the correlation matrix, we obtain a fully connected network for which $C=\hat{C}=1$ for all times, whereas $\tilde{C}$ still shows the effect of the crash clearly.

(B) *Directed metabolic network*. Cellular metabolism can be represented as a directed network of intracellular molecular interactions. The network consists of nodes X_i, Y_j, which represent the chemicals and they are linked if connected by a metabolic reaction. Here we focus on the metabolic pathways of the bacterium *Escherichia coli* grown in glucose, which has been studied intensely [4]. In order to experiment with weighted directed motifs, we define the weights through a biochemical reaction of the form $x_1X_1+\cdots x_nX_n \to y_1Y_1+\cdots y_mY_m$ with a positive (negative) net flux f if the balance of the reaction lies to the right (left). The flux provides an overall measure of the relative activity of each reaction. We define the weights as $w_{ij}=(y_j/x_i)f$, reflecting the rate at which X_i is converted into Y_j.

In order to employ motif intensity scores a reference system, corresponding to a null hypothesis, needs to be established. We follow a typical approach by constructing an ensemble of random networks by conserving the degree sequence of the empirical network using a switching algorithm [13], which preserves the single-node characteristics of the empirical network. The weights are obtained simply by permuting the empirical weights. While removing any

weight correlations, the approach guarantees conservation of the empirical weight distribution.

We summarize our findings in Fig. 3, in which we show the unweighted and weighted motif intensities for a subset of the studied motifs: (i) path of order 2, (ii) nonfrustrated triangle, and (iii) frustrated triangle. The motif intensity scores for the unweighted networks, which are based on the subgraph counts, are $z_i=-5.4$, $z_{ii}=12.8$, and $z_{iii}=-0.5$, and for the weighted networks $\tilde{z}_i=14.8$, $\tilde{z}_{ii}=33.8$, $\tilde{z}_{iii}=9.0$. These results show that a move from unweighted to weighted characteristics can cause a change from low to high intensity, i.e., from under-representation to over-representation. The intensity may also become amplified, or change from statistically insignificant to being over-represented.

In this paper we have proposed two concepts for the characterization of weighted complex networks: the intensity and coherence of a subgraph. They allow for a very natural generalization of the z scores to motif intensity scores [Eqs. (6) and (7)], and the clustering coefficient to weighted clustering coefficient [Eq. (10)]. Our studies with undirected financial networks show that the weighted clustering coefficient reflects the effects of a market crash which is hardly observed with other clustering characteristics studied. Our results on the directed metabolic network of E. Coli indicate that incorporation of weights into network motifs may considerably modify the conclusions drawn from their statistics.

We are thankful to A.-L. Barabási, E. Almaas, and S. Wuchty for the metabolic network data and useful discussions. This work was carried out at the Center of Excellence of the Finnish Academy of Sciences, Computational Engineering, HUT. J.K. was partially supported by the Center for Applied Mathematics and Computational Physics, BUTE.

[1] R. Albert and A.-L. Barabási, Rev. Mod. Phys. **74**, 47 (2002); S. N. Dorogovtsev and J. F. F. Mendes, Adv. Phys. **51**, 1079 (2002); M. E. J. Newman, SIAM Rev. **45**, 167 (2003).

[2] R. Pastor-Satorras and A. Vespignani, *Evolution and Structure of the Internet: A Statistical Physics Approach*, edited by R. R. Guimera, S. Mossa, A. Turtschi, and L. A. N. Amaral (Cambridge University Press, Cambridge, 2004).

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

[4] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A.-L. Barabási, Nature (London) **407**, 651 (2000); T. R. Hughes *et al.*, Cell **102**, 109 (2000); S. Light and P. Kraulis, Bioinformatics **5**, 15 (2004); O. Fiehn and W. Weckwerth, Eur. J. Biochem. **270**, 579 (2003); E. Almaas, B. Kovács, T. Vicsek, Z. N. Oltvai, and A.-L. Barabási, Nature (London) **427**, 839 (2004).

[5] D. K. Slonim, Nat. Genet. **32**, 502 (2002); I. J. Farkas, H. Jeong, T. Vicsek, A.-L. Barabási, and Z. N. Oltvai, Physica A **318**, 601 (2003); K. Rho, H. Jeong, and B. Kahng, cond-mat/0301110; A. Vazquez, A. Flammini, A. Maritan, and A. Vespignani, Nat. Biotechnol. **21**, 697 (2003).

[6] R. N. Mantegna, Eur. Phys. J. B **11**, 193 (1999); L. Kullmann, J. Kertész, and R. Mantegna, Physica A **287**, 412 (2000); L. Giada and M. Marsili, Phys. Rev. E **63**, 061101 (2001).

[7] J.-P. Onnela, K. Kaski, and J. Kertész, Eur. Phys. J. B **30**, 285 (2004); J.-P. Onnela, K. Kaski, and J. Kertesz, *ibid.* **38**, 353 (2004).

[8] L. A. Braunstein, S. V. Buldyrev, R. Cohen, S. Havlin, and H. E. Stanley, Phys. Rev. Lett. **91**, 168701 (2003).

[9] G. Szabó, M. J. Alava, and J. Kertész, Physica A **330**, 31 (2003); P. J. Macdonald, E. Almaas, and A.-L. Bararabási, cond-mat/0405688.

[10] T. Schank and D. Wagner (unpublished).

[11] A. Barrat, M. Barthélemy, and A. Vespignani, Phys. Rev. Lett. **92**, 228701 (2004); cond-mat/0406238.

[12] M. E. J. Newman, cond-mat/0407503.

[13] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, Science **298**, 824 (2002); S. Wuchty, Z. N. Oltvai, and A.-L. Barabási, Nat. Genet. **35**, 176 (2003); R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, and U. Alon, Science **303**, 1538 (2004).

[14] Y. Artzy-Randrup, S. J. Fleishman, N. Ben-Tal, and L. Stone, Science **305**, 1107c (2004); R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, and U. Alon, *ibid.* **305**, 1107d (2004).

[15] G. Szabó, M. J. Alava, and J. Kertész, in *Complex Networks*, edited by Eli Ben-Naim *et al.*, Springer Lecture Notes in Physics No. 650 (Springer, Berlin, 2004), pp. 139–162.

[16] J.-P. Onnela, A. Chakraborti, K. Kaski, and J. Kertész, Physica A **324**, 247 (2003).