© 2005 American Institute of Physics. Reprinted with permission from J. F. F. Mendes, et al. (Eds.), Science of Complex Networks, AIP Conference Proceedings, 776, 108 (2005).

Characterizing Motifs in Weighted Complex Networks

Jari Saramäki[*], Jukka-Pekka Onnela[*], Janos Kertész[†,*] and Kimmo Kaski[*]

[*]*Laboratory of Computational Engineering, Helsinki University of Technology, Espoo, Finland*
[†]*Department of Theoretical Physics, Budapest University of Technology and Economics, Budapest, Hungary*

Abstract.
The local structure of unweighted complex networks can be characterized by the occurrence frequencies of subgraphs in the network. Frequently occurring subgraphs, *motifs*, have been related to the functionality of many natural and man-made networks. Here, we generalize this approach for *weighted* networks, presenting two novel measures: the *intensity* of a subgraph, defined as the geometric mean of its link weights, and the *coherence*, depicting the homogeneity of the weights. The concept of motif scores is then generalized to weighted networks using these measures. We also present a definition for the weighted clustering coefficient, which emerges naturally from the proposed framework. Finally, we demonstrate the concepts by applying them to financial and metabolic networks.

Keywords: Weighted complex networks, motifs, clustering coefficient
PACS: 89.75.-k, 89.75.Hc, 89.65.-s, 87.16.Ac

INTRODUCTION

The network approach to complex systems has turned out to be extremely fruitful during the last years, revealing some general principles applicable to a large number of systems, ranging from the Internet to the protein-protein interaction networks of living cells [1, 2, 3]. The main strength of this approach is its ability to capture essential features of the systems in question by using simple building blocks, the vertices and edges, which represent elements or units and their interactions. Studies of the characteristics of networks have produced novel, unexpected findings such as the very small average shortest vertex-to-vertex distances often encountered in natural networks, the ubiquity of scale-free topologies, often coupled with high clustering and some signatures of modularity, as well as the significantly high frequency of specific network motifs [4, 5, 6], which can be considered as basic structural elements of networks.

In order to understand better the complex systems under study, it is evident that information about the nature and strength of the underlying interactions should be taken into account. A natural way of doing this is to assign weights to the network edges, such as those provided by fluxes in transportation-related networks, e.g. the Internet and air traffic networks [7, 8, 9], or fluxes of chemical species produced in reactions like those building the metabolic pathways of a cell [10, 11, 12]. Yet another way to obtain a weighted network is to utilize correlation matrices to identify the system structure, e.g., for inferring the underlying dynamics of stock market data [13, 14, 15].

The above examples indicate the need to generalize commonly utilized network

characteristics to weighted networks. Here, we will focus on measures of weighted *subgraphs* and present an extended discussion of the concepts introduced in [16]. The goal is to provide practical tools for characterizing the importance of specific *motifs*, i.e., frequently occurring topologically equivalent subgraphs in weighted networks.

INTENSITY AND COHERENCE OF MOTIFS

Motifs with significantly frequent occurence have been related to functional properties of, e.g., biological and social networks. In the unweighted case, the standard approach involves counting the number of times a specific type of subgraph appears in a network, and comparing the appearance frequency to a randomized reference ensemble. However, in the weighted framework, information would be lost by taking only the appearance frequency of subgraphs into account. Further, we may consider any weighted network as a fully connected graph where some links bear zero weights. Counting the number of times a specific subgraph appears would require imposing a threshold condition on its weights - should a subgraph where one link bears a vanishing weight ε be included in the count or not?

To overcome the above-mentioned difficulties, we define the *intensity* $I(g)$ of subgraph g with vertices v_g and links ℓ_g as the *geometric mean* of its weights:

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

where $|\ell_g|$ is the number of links in ℓ_g, and w_{ij} denotes the weight of the link between vertices i and j. The weights are considered to be non-negative, but not necessarily normalized. Evidently, as the edge weights are multiplied, the intensity is zero if any of the weights is zero, and becomes small if any of the weights is small. The definition suggests a shift in perspective from regarding subgraphs as discrete objects, which 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 (see Fig. 1).

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" [4]. This has led to some confusion, which partly stems from the specification of the random ensemble, i.e. the underlying null hypothesis [17, 18]. Further, the terms "subgraph" and "motif" have been used interchangeably by some authors [5]. For the sake of clarity, we choose here to define *a motif as a set (an ensemble) of topologically equivalent subgraphs of a network*. With weighted networks it then becomes more natural to deal with motif intensities as opposed to numbers of occurrence; the latter is obtained as a special case of the former when the weights are considered binary. The motifs showing statistically significant deviation from some reference system can then be called high or low intensity motifs.

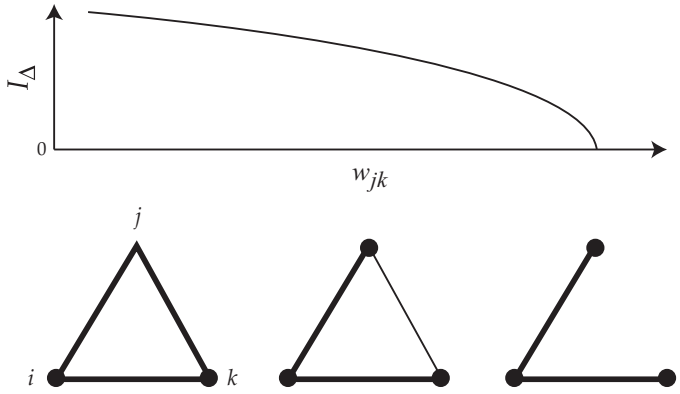

FIGURE 1. A schematic illustration of the intensity of a triangle motif, when the weight w_{jk} of one of the constituent edges is gradually decreased from left to right. Here, the intensity $I_\Delta \propto w_{jk}^{1/3}$.

In this framework, the *total intensity* I_M of a motif M in the network can now be defined as the sum of its subgraph intensities:

$$I_M = \sum_{g \in M} I(g). \tag{2}$$

To examine the significance of the total intensity of a motif in an empirical network, this quantity should be compared to a reference system. For unweighted networks, the statistical significance of motif occurrence is indicated by the z-score, defined as

$$z_M = (N_M - \langle n_M \rangle)/\sigma_M, \tag{3}$$

where N_M is the number of subgraphs in motif M in the empirical network, $\langle n_M \rangle$ is the expectation of their number in the reference ensemble, and σ_M is the standard deviation of the latter. This concept is readily generalized to weighted motifs by replacing the number of subgraphs by their intensities, and we may now define the *motif intensity score* as

$$\tilde{z}_M = (I_M - \langle i_M \rangle)/\sqrt{\langle i_M^2 \rangle - \langle i_M \rangle^2}, \tag{4}$$

where i_M is the total intensity of motif M in one realization of the reference system[1]. It is clear that Eqs. (3) and (4) coincide for binary weights, implying that $\tilde{z} \to z$ in the limit.

However, due to the nature of the geometric mean, the intensity $I(g)$ does not bear information on the weight distribution inside a subgraph; it 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

[1] In the unweighted case, the common approach for constructing the reference system, i.e. the underlying null hypothesis against which statistical significance is tested, is to generate an ensemble of random networks such that the degree sequence of the empirical network is conserved in the randomized networks.

ratio of the geometric to the arithmetic mean of the weights as

$$Q(g) = I(g)|\ell_g| / \sum_{(ij)\in \ell_g} w_{ij}. \tag{5}$$

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. Analoguously to the motif intensity score, we can also define the *motif coherence score* as

$$\widetilde{z'}_M = (Q_M - \langle q_M \rangle) / \sqrt{\langle q_M^2 \rangle - \langle q_M \rangle^2}, \tag{6}$$

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.

THE WEIGHTED CLUSTERING COEFFICIENT

As triangles are one type of subgraph, it is of interest to consider generalizing the *clustering coefficient* C to weighted networks in the present framework. For unweighted networks, the clustering coefficient at vertex i of degree k_i is defined as

$$C_i = \frac{2t_i}{k_i(k_i-1)}, \tag{7}$$

i.e., it is the ratio of the number t_i of triangles where vertex i participates to the maximum possible number of such triangles. Hence, $C_i \in [0,1]$. There is no single, evident way to generalize this concept to the weighted case, and several proposals exist [9, 16, 19]. Our version of the weighted clustering coefficient $\widetilde{C}$, introduced in [16], is based on the following additional requirements:

1. As the weights become binary, $\widetilde{C} = C$.
2. For compatibility, $\widetilde{C} \in [0,1]$.
3. In the unweighted case, the number of triangles at a node determines its clustering properties. In the weighted case, clustering should be determined by some weighted characteristic of triangles.
4. For each triangle, all three edge weights should be taken into account.
5. For each triangle, the weighted characteristic should be invariant to permutation of weights[2].
6. When any of the weights in a triangle approaches zero, the weighted characteristic of that triangle should likewise approach zero.
7. When vertex i participates in the maximum number $\frac{1}{2}k_i(k_i-1)$ of triangles, where each edge weight is maximal, the weighted clustering coefficient should also be maximal, i.e., $\widetilde{C}_i = 1$.

[2] This ensures that for a single triangle ijk, the value of the weighted clustering coefficient is the same at all vertices, i.e., $\widetilde{C}_i = \widetilde{C}_j = \widetilde{C}_k$

These requirements can be fulfilled by replacing the number of triangles in Eq. (7) with the sum of triangle intensities. Then, the weighted clustering coefficient can be defined as

$$\widetilde{C}_i = \frac{2}{k_i(k_i-1)} \sum_{j,k} (\tilde{w}_{ij}\tilde{w}_{jk}\tilde{w}_{ki})^{1/3}, \tag{8}$$

where the weights are scaled by the largest weight in the network, $\tilde{w}_{ij} = w_{ij}/W$ with $W = \max(w_{mn})$. This normalization ensures that requirements (2) and (7) are fulfilled. Note that normalization based on *weighted single-node characteristics* such as the node strength $s_i = \sum_j w_{ij}$ would violate the weight permutation invariance requirement (5). Further, we can see that requirements (4)-(7) are also fulfilled by the unweighted clustering coefficient C.

We can relate the weighted clustering coefficient to the unweighted one through the *average intensity* of triangles at vertex i as $\overline{I}_i = \frac{1}{t_i} \sum_{g_\Delta \in \mathscr{N}(i)} I(g_\Delta)$, where $\mathscr{N}(i)$ denotes the neighborhood of node i. This allows us to write the weighted clustering coefficient as

$$\widetilde{C}_i = \overline{I}_i C_i. \tag{9}$$

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 (taken with the normalized weights).

APPLICATION I: CLUSTERING IN AN UNDIRECTED FINANCIAL NETWORK

We have applied the proposed weighted clustering coefficient to the analysis of a financial interaction network inferred from a set of daily price data for $N = 477$ NYSE traded stocks from the years 1980 to 2000. We first calculated correlation matrices by extracting sliding 4-year return windows in order to study the system's dynamics. Then, for each window, we constructed a network depicting the financial interactions, such that its vertices correspond to stocks, and the weighted undirected edges to the elements in the corresponding correlation matrix. To be precise, the weights are taken as the absolute values of the correlation coefficient. Hence, strong edge weights imply strong coupling between the stock returns in terms of their linear correlation. Evidently, including every correlation matrix element would result in a fully connected network – to avoid this, edges have been included in the network in descending order of weight, until a predetermined number of links has been reached (in the example discussed below, 476 links were used). For a detailed description of the method, see [14, 15].

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 [20], such that during the crash the network shrinks and correlations increase. Our aim is to use it as an example of a network undergoing this type of two-fold transition – topology and weights – and to see whether the changes are reflected in the network's clustering properties. For comparison, we have used the unweighted clustering coefficient C of Eq. (7), and an alternative weighted coefficient $\hat{C}$ defined in [9] as

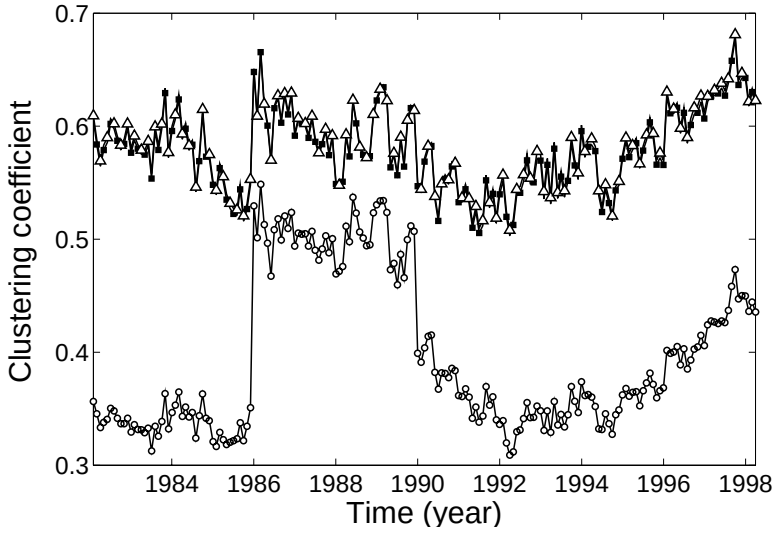

FIGURE 2. Average clustering coefficients for the financial network. The weighted clustering coefficient $\widetilde{C}$ (○) of Eq. (8) shows the effect of Black Monday clearly. The unweighted C (■) of Eq. (7) and the weighted $\hat{C}$ (△) of Eq. (10) practically coincide (the markers ■ and △ are used alternately).

$$\hat{C}_i = \frac{1}{s_i(k_i-1)} \sum_{j,k} \frac{(w_{ij}+w_{ik})}{2} a_{ij} a_{ik} a_{jk}, \tag{10}$$

where s_i denotes the strength of node i, defined as $s_i = \sum_j w_{ij}$, 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.

Fig. 2 displays values of all three clustering coefficients as functions of time, averaged over all vertices. The changes of the unweighted coefficient C do not capture the effect of the crash very clearly, as its value is solely determined by the topological aspects of the transition. The weighted coefficient $\hat{C}$ also appears fairly insensitive to the changes in link weights in these networks, and its values practically coincide with the unweighted C. However, the weighted $\widetilde{C}$ is clearly seen to reflect the transition, indicating its ability to capture both aspects of the transition. The time-averaged values for the clustering coefficients outside (inside) the crash period are $C_{avg} = 0.57$ ($C_{avg} = 0.60$), $\hat{C}_{avg} = 0.58$ ($\hat{C}_{avg} = 0.60$), and $\widetilde{C}_{avg} = 0.36$ ($\widetilde{C}_{avg} = 0.50$). These numbers imply that C_{avg} and $\hat{C}_{avg}$ increase less than 5% during the crash which is less than the normal fluctuation of C and $\hat{C}$ outside the crash period, measured at 6.2% as their standard deviation relative to the mean. However, the crash increases $\widetilde{C}_{avg}$ by 39%, which is considerably larger than the the level of fluctuation at 9.7%. Thus, $\widetilde{C}$ has a considerably higher "signal-to-noise" ratio. Further, changing the value of the manually set weight threshold, which determines the number of edges included in the financial interaction network, does not affect the results in a significant way. Even in the limit of including every element of the correlation matrix as an edge, which results in a fully connected network for which

the unweighted clustering coefficient $C = \hat{C} = 1$ for all times, the weighted clustering coefficient $\widetilde{C}$ was still seen capture the effect of the crash clearly.

APPLICATION II: MOTIFS IN A DIRECTED METABOLIC NETWORK

The concept of motifs is especially relevant in the case of biological networks, such as those depicting gene regulation, protein interactions, or metabolism, where subgraph connectivity can often be related to functional "modules". These types of biological networks are well suited for the weighted framework, as in most cases there is a natural way for including the interaction strengths or, in the case of metabolic networks, reaction fluxes, into the network depiction as edge weights. Cellular metabolism can be represented as a directed network of intracellular molecular interactions, such that the network consists of vertices X_i, Y_j, which represent the chemicals. Then the vertices are connected by an edge if the chemicals are connected by a metabolic reaction. Here, we have focused on the metabolic pathways of the bacterium *Escherichia coli* grown in glucose, which have been studied intensely (see, e.g., [10, 21]), and analyzed the intensities of weighted motifs in the related network. Here, our aim has been merely to point out that once weights are considered, findings may strongly differ from the unweighted case. We have chosen not yet to consider possible biological interpretations of these findings.

In order to experiment with weighted directed motifs, we define the weights through a biochemical reaction of the form $x_1X_1 + \cdots x_nX_n \rightarrow y_1Y_1 + \cdots y_mY_m$ 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, allowing us to define the corresponding weights as $w_{ij} = (y_j/x_i)f$, reflecting the rate at which X_i is converted into Y_j. In addition, for analyzing motif significances based on intensity scores, a reference system needs to be established, corresponding to a null hypothesis. We follow a typical approach by computationally constructing an ensemble of random networks by conserving the degree sequence of the empirical network using a switching algorithm [6], which preserves the single-node characteristics of the empirical network. Likewise, the weight distribution is conserved by simply permuting the edge weights, which removes any weight correlations.

Our findings are displayed in Fig. 3, where the unweighted and weighted motif intensities are shown for a subset of the studied motifs: (i) path of order 2, (ii) non-frustrated triangle, and (iii) frustrated triangle [3]. The unweighted subgraph-count-based z-scores (the weighted intensity-based $\tilde{z}$-scores) are $z_i = -5.4$ ($\tilde{z}_i = 14.8$), $z_{ii} = 12.8$ ($\tilde{z}_{ii} = 33.8$), and $z_{iii} = -0.5$ ($\tilde{z}_{iii} = 9.0$). These results indicate that a shift from unweighted to

[3] Here, for compatibility with earlier work, subgraphs have been counted only once, that is, a closed triangle prevents the counting of the three open triangles (paths of order 2) it contains. However, there are evident problems with this approach; following this logic, one should specify an arbitrary upper limit as to what is counted. Especially in cases where noise might be manifested as extra edges with small weights, this approach may lead to erroneous conclusions as "true" motifs are not counted, whereas larger motifs that are merely an artefact due to noise are.

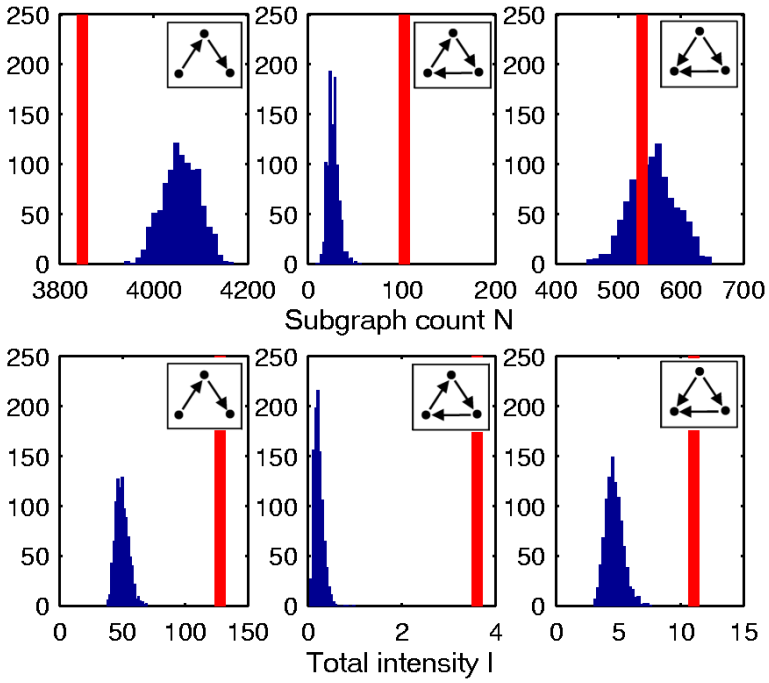

FIGURE 3. Motif intensities for the empirical network (vertical lines) and the corresponding random ensembles (histograms), for the unweighted (upper panel) and weighted (lower panel) cases.

weighted characteristics can cause a change from low to high intensity, i.e. from under-representation to over-representation, as in case (i). The intensity may become amplified, i.e. increase the extent of over-representation (case (ii)), or it may increase from average to high intensity, i.e. from statistically insignificant to over-representation (case (iii)). Thus, we argue that when investigating the significance of motifs in biological networks, the inclusion of weights in the analysis can dramatically change the picture.

It is also worth noting that the proposed framework is especially suitable for situations with experimental errors and noise, which may be the case when dealing with biological networks inferred from measurement data. As the subgraph intensities form a continuum, the effect of noise is also "continuous" and less drastic than when the subgraphs are treated in a binary way by imposing a threshold weight. In principle, the intensity scores of motifs should still be reasonably reliable even in the presence of "background noise" resulting in a large number of extra edges with small weights.

SUMMARY AND CONCLUSIONS

To summarize, we have discussed novel concepts for the characterization of subgraphs and motifs in weighted complex networks: subgraph *intensity* and *coherence*, as well as the *weighted clustering coefficient* which emerges directly within the proposed framework. These measures suggest a shift in perspective from subgraphs being binary objects, which either exist or not, to a continuum of subgraph intensities, and allow a

natural generalization of the z-scores measuring motif significance. We have applied these concepts to two cases, undirected financial and directed metabolic networks. We have shown that our version of the weighted clustering coefficient clearly captures the effects of a market crash, and that the incorporation of weights into the network motifs of the metabolic network of *E. Coli* considerably modifies their statistics. We hope that the work presented herein will stimulate generalizing existing network concepts to the weighted framework, and that the proposed characteristic measures will find further use in future studies of weighted complex networks.

ACKNOWLEDGMENTS

We are thankful to A.-L. Barabási, E. Almaas and S. Wuchty for the *Escherichia Coli* metabolic network data and useful discussions. This work was carried out at the Center of Excellence of the Finnish Academy of Sciences, Computational Engineering, Helsinki University of Technology. JK is partially supported by the Center for Applied Mathematics and Computational Physics, BUTE.

REFERENCES

1. R. Albert, and A.-L. Barabási, *Rev. Mod. Phys.*, **74**, 47 (2002).
2. S. Dorogovtsev, and J. Mendes, *Adv.Phys*, **51**, 1079–1187 (2002).
3. M. Newman, *SIAM Review*, **45**, 167–256 (2003).
4. R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, *Science*, **298**, 824–827 (2002).
5. S. Wuchty, Z. Oltvai, and A.-L. Barabási, *Nature Genetics*, **35**, 176–179 (2003).
6. R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, and U. Alon, *Science*, **303**, 1538–1542 (2004).
7. R. Pastor-Satorras, and A. Vespignani, *Evolution and Structure of the Internet: A Statistical Physics Approach*, Cambridge University Press, 2004.
8. R. Guimera, S. Mossa, A. Turtschi, and L. Amaral (2003), `cond-mat/0312535`.
9. A. Barrat, M. Barthélemy, R. Pastor-Satorras, and A. Vespignani, *Proc. Natl. Acad. Sci USA*, **101**, 3747 (2004).
10. H. Jeong, B. Tombor, Z. Oltvai, and A.-L. Barabási, *Nature*, **407**, 651 (2000).
11. J. Stelling, S. Klamt, K. Bettenbrock, S. Schuster, and E. Gilles, *Nature*, **420**, 190 (2002).
12. E. Almaas, B. Kovács, T. Vicsek, Z. Oltvai, and A.-L. Barabási, *Nature*, **427**, 839 (2004).
13. R. Mantegna, *Eur. Phys. J. B*, **11**, 193 (2000).
14. J.-P. Onnela, A. Chakraborti, K. Kaski, and J. Kertész, *Eur. Phys. J. B*, **30**, 285–288 (2002).
15. J.-P. Onnela, K. Kaski, and J. Kertész, *Eur. Phys. J. B*, **38**, 353–362 (2004).
16. J. Onnela, J. Saramäki, J. Kertész, and K. Kaski (2004), `cond-mat/0408629`.
17. Y. Artzy-Randrup, S. Fleishman, N. Ben-Tal, and L. Stone, *Science*, **305**, 1107c (2004).
18. R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, and U. Alon, *Science*, **305**, 1107d (2004).
19. P. Holme, S. Park, B. Kim, and C. Edling (2004), `cond-mat/0411634`.
20. J.-P. Onnela, A. Chakraborti, K. Kaski, and J. Kertész, *Physica A*, **324**, 247 (2003).
21. S. Light, and P. Kraulis, *BMC Bioinformatics*, **5**, 15 (2004).