Offdiagonal Complexity: A computationally quick complexity measure for graphs and networks

(Physica A, in print)

Jens Christian Claussen

Institut für Theoretische Physik und Astrophysik,
Universität Kiel, Leibnizstraße 15, 24098 Kiel, Germany
phone ++49-431-880-4096
claussen@theo-physik.uni-kiel.de

Abstract

A vast variety of biological, social, and economical networks shows topologies drastically differing from random graphs; yet the quantitative characterization remains unsatisfactory from a conceptual point of view. Motivated from the discussion of small scale-free networks, a biased link distribution entropy is defined, which takes an extremum for a power law distribution. This approach is extended to the node-node link cross-distribution, whose nondiagonal elements characterize the graph structure beyond link distribution, cluster coefficient and average path length. From here a simple (and computationally cheap) complexity measure can be defined. This Offdiagonal Complexity (OdC) is proposed as a novel measure to characterize the complexity of an undirected graph, or network. While both for regular lattices and fully connected networks OdC is zero, it takes a moderately low value for a random graph and shows high values for apparently complex structures as scale-free networks and hierarchical trees. The Offdiagonal Complexity apporach is applied to the Helicobacter pylori protein interaction network and randomly rewired surrogates.

1 Introduction

While random graph theory and scale-free network research know a set of standard measures to quantify their properties, the question of *complexity* of a graph still is in its infancies. A 'blind' application of other complexity

Preprint submitted to Elsevier Science October 16, 2004; revised August 11, 2006

measures (as for binary sequences or computer programs) does not account for the special properties shared by graphs and especially scale-free graphs. Moreover, some known complexity measures themselves have a high computational complexity.
Since a series of seminal papers (Watts & Strogatz [1], Barabasi & Albert [2] [2,3], Newman [4], Dorogovtsev & Mendes [5]) since 1999 (see also [6] for an overview), small-world and scale-free networks are a hot topic of investigation in a broad range of systems and disciplines. Metabolic and other biological networks, collaboration networks, www, internet, etc., have in common that the distribution of link degrees follows a power law, thus has no inherent scale. Such networks are termed 'scale-free networks'. Compared to random graphs, which have a Poisson link distribution and thus a characteristic scale, they share a lot of different properties, especially a high clustering coefficient, and a short average path length.
Mathematically, a graph (or synonymously in this context, a network) is defined by a (nonempty) set of nodes, a set of edges (or links), and a map that assigns two nodes (the "end nodes" of a link) to each link. In a computer, a graph may be represented either by a list of links, represented by the pairs of nodes, or equivalently, by its adjacency matrix a_{ij} whose entries are 1 (0) if nodes i, j are connected (disconnected). Useful generalizations are weighted graphs, where the restriction of a_{ij} is relaxed from binary values to (unsually nonnegative) integer or real values (e.g. resistor values, travel distances, interaction coupling), and directed graphs, where a_{ij} no longer needs to be symmetric, and the link from i to j and the link from j to i can exist independently (e.g. links between webpages, or scientific citations).
Here the discussion will be kept limited to binary undirected graphs, like an acquaintancy network or a railway network as shown below. In the following sections the link (degree) distribution and the next order cross-distribution are investigated and taken as a basis for a complexity measure.

2 Other complexity measures

For text strings (as computer programs, or DNA) there are common complexity measures in theoretical computer science, as *Kolmogorov complexity* (and the related *Lempel-Ziv complexity* and *algorithmic information content* AIC) [8]. E.g., AIC is defined by the length of the shortest program generating the string. For random structures, thus also for random graphs, they indicate high complexity. A distinction of complex structured (but still partly random) structures from completely random ones usually is prohibitive for this class of measures. For this reason, measures of *effective complexity* [9] have been discussed; usually these are defined as an entropy (or description length) of

"a concise description of a set of the entity's regularities" [9]. Here we are mainly interested in this second class, and straightforwardly one would try to apply existing measures, e.g., to the link list or to the adjacency matrix. However, mathematically it is not straightforward to apply these text string based measures to graphs, as there is no unique way to map a graph onto a text string. For the case of hierarchical structures, which can be represented by trees, Ceccatto and Huberman quantified complexity from the diversity of the subtrees [7]. As natural networks typically exhibit an occurrence triangles and higher oder loops in a nonneglectable way, other approaches have to be chosen for networks in general.

Thus one desires to use complexity measures that are defined directly for graphs. Two classical measures are known from graph theory, *graph thickness* and *coloring number* have a low "resolution" (typically integer values up to 4), and their relevance for real networks is not clear. Two new complexity measures recently have been proposed for graphs, *Medium Articulation* [10] for weighted graphs (as they appear in foodwebs) and a measure for directed graphs by Meyer-Ortmanns [11] based on the *network motif* concept [12]). Unfortunately, the latter two complexity measures are computationally quite costly. A computational complexity approach has been defined by Machta and Machta [13] as *computational depth* of an *ensemble of graphs* (e.g. small-world, scalefree, lattice). It is defined as the number of processing time steps a large parallel computer (with unlimited number of processors) would need to generate a *representative* member of that graph ensemble. Unlike other approaches, it does not assign single complexity values to each graph, and again is nontrivial to compute.

Following [9], an especially desired property of a complexity measure should be the ability to distinguish nonrandom complex structures from both pure randomness and regular structures as lattices. In this instance, the effective complexity and the Machta approach fulfill this prerequisites perfectly, but up to today no simple method is available to compute them. Hence, a *simpler estimator* of graph complexity is desired, and one possible approach, the Offdiagonal Complexity, is proposed here. It is motivated by a striking observation on the node-node link correlation matrices of complex networks [14], namely that entries are more evenly spread among the offdiagonals, compared to both regular lattices and random graphs (see Figs. 4 and 5 for a comparison). This can be used to define a complexity measure, for undirected graphs [14,15].

This article is organized as follows. In Sec. 3 the approach is motivated from link entropies and node-node correlations. In Sec. 4 OdC is defined. Section 5 investigates the application of OdC to a protein interaction network, compared with randomized surrogates.

3 Motivation of OdC

3.1 Node degree correlations: Methods of classical statistics

A straightforward mathematical approach to study node-node link correlations, i.e. correlations between degrees of pairs of nodes, is to use rank correlation methods [16] from classical statistics to analyze the link distributions.

Two common rank correlation methods can be described as follows. One considers a list of *rank numbers* of link numbers (node degrees). For each of the two graphs (A and B) to be compared, there is a (ordered) list of link numbers $(k_1, k_2, \cdots k_N) = 5\ 2\ 2\ 1\ 1\ 1$, and one assigns a rank number to each link, $(r_1^{\mathrm{A}}, r_1^{\mathrm{A}}, \ldots r_N^{\mathrm{A}}) = (1\ 2.5\ 2.5\ 5\ 5\ 5)$. Hereby the identical second and third ranks are replaced by the (noninteger) average value; as node degrees are highly degenerate, this will occur frequently.

Then the Kendall tau coefficient is defined as $t = \dfrac{2\sum_{ij}\sigma_{ij}}{n(n-1)}$, where $\sigma_{ij} = \pm 1$, if pairs of elements (i, j) are ranked in both lists equally (resp. non equally), $\sigma_{ij} = \mathrm{sgn}(r_i^{\mathrm{A}} - r_j^{\mathrm{A}}) \cdot \mathrm{sgn}(r_i^{\mathrm{B}} - r_j^{\mathrm{B}})$. Its apparent drawbacks within this context are the required costly computations (n^2), and it seems to be analytically not easy to handle, as one must have the nodes sorted by their degree, for each member of (e.g.) an ensemble average.

The second main rank correlation method is Spearman's rho, defined by $r_s = 1 - \dfrac{6\sum_i d_i^2}{n^3 - n}$, where $d_i = r_i^{\mathrm{A}} - r_i^{\mathrm{B}}$. — Some of its properties are:

$r_s = +1$ for *identical* rank lists $\begin{matrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{matrix}$

$r_s = -1$ for *counter-sequenced* rank lists $\begin{matrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{matrix}$

$r_s = +\frac{1}{2}$ if a sequence is constant $= \frac{n(n-1)}{2}$. (One might wonder why not $r_s = 0$ holds here. However for $n = 3$ always $r_s \neq 0$ holds; but the average over all possible rank lists vanishes, $\langle r_s \rangle = 0$.)

In general, rank correlation methods are *not appropriate* for a *high degeneracy*, i.e. a large number of nodes with the same number of links.

Thus, it is desired to formulate other measures that can estimate the complexity of a graph from correlation information of pairs of nodes. The approach of this paper is to define an entropy-type measure. To motivate the ansatz, the problem of binning and the definition of a link entropy is discussed in the next section.

3.2 Fit of sparse power-law distributions

The fit of sparse distributions by binning has to cope with the problem of zeroes and with the effect of arbitrariness of the choice of interval length and position. As an example we consider the link distribution of a traffic network [17] (see Fig. 1).

As intervals have to be chosen so that no zeroes occur ($-\infty$ in log scale), one has the choice between different 'tricks' (influencing the fit): (i) irregular intervals: choice influences fit, or (ii) regular intervals $n_{\max} \cdot \sqrt{2}\ln(c \cdot \exp(k))$, however they imply a severe reduction of the number of intervals. Even the two remaining parameters influence the result (esp. for large link numbers): (see Fig. 2b). A moderately 'clean' method is to place the entry with largest link number in the middle of that interval. A parameter-free approach is the integrated density. For a power law density with exponent $\alpha > 1$, one has

$$\int_x^\infty \mathrm{d}k \;\; k^{-\alpha} = \frac{x^{1-\alpha}}{\alpha - 1},$$

Instead of the density itself, the integrated density can be fitted (see Fig. 2c). For exact results, a discretization correction c_n^α is necessary: $c_n^\alpha = \frac{\sum_{k=n}^\infty k^\alpha}{\int_n^\infty \mathrm{d}k \;\; k^{-\alpha}}$. Alternatively, from $\sum_{k=n}^\infty k \cdot p(k)$ one gets a plot with the same slope as $p(k)$ itself.

3.3 Entropy of the link distribution

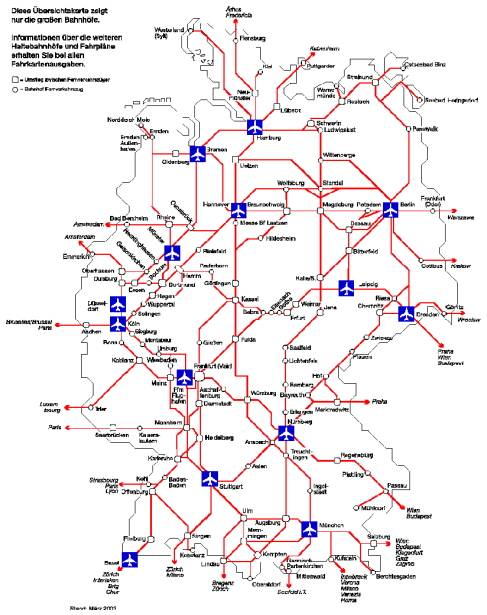
As demonstrated in sec. 3.2, the estimation of the scale exponent from a measured distribution by binning has inherent degrees of freedom; this can be overcome by a fit of the integrated density. To estimate the entropy of a distribution ($\neq$ density) with sampling gaps however leads to underestimation (Grassberger [18]). A straightforwardly defined link distribution entropy $H = -\sum_k p_k \ln p_k$ becomes extremal for the *equidistribution* (and not for a power law). Power law candidate distribution are usually logarithmically binned. However, for a power law one obtains a distribution with linear decay (in the binned log-log space, as in Fig. 2b,c), and not an equidistribution, and again H not maximal.

This problem is solved by defining a "Biased Link Entropy" (showing an extremum w.r.t. α, see Fig. 3; the transformed density is the equidistribution for proper choice of α). With the necessary normalization $N(\alpha) = \sum_k k^\alpha p_k/\delta_k$, here δ_k may be a binning interval width, the biased link entropy reads

$$H(\alpha) = -\sum_k \frac{k^\alpha p_k}{\delta_k N(\alpha)} \ln \frac{k^\alpha p_k}{\delta_k N(\alpha)}. \tag{1}$$

3.4 Node degree correlations: Entropy approach?

The idea now is to use entropies instead of correlations or rank correlations. Naively one would use define an entropy of all coefficients of the node degree correlation matrix p_{kl}, $H = -\sum_{kl} p_{kl} \ln p_{kl}$. However, then any invariances like $(k_1, k_2) \to (2k_1, 2k_2)$ or $(k_1, k_2) \to (k_0 + k_1, k_0 + k_2)$ are lost, but such invariances would be desired for different description levels of the systems. Annother possible approach could be via the Kullback-Leibler Distance $D(p^{\mathrm{A}}, p^{\mathrm{B}}) = \sum_i p_i^{\mathrm{A}} \ln(p_i^{\mathrm{A}}/p_i^{\mathrm{B}})$. Here, one has to apply it to the node degree $k_i^{\mathrm{A}}, k_i^{\mathrm{B}}$ for each link i. However, this is generically nonsymmetric (for a symmetrized definition see [19]), and again, there is no invariance for e.g. $(k_1, k_2) \to (2k_1, 2k_2)$. — As a last approach, one could define a Biased Cross Link Entropy by replacing k_i^{B} by $k^\gamma \cdot k_i^{\mathrm{B}}$. — This discussion shows that simple definitions via link entropies bear difficulties.



Node degree	# of nodes
3	20
4	11
5	6
6	2
7	1
8	0
9	0
10	1
11	0
12	0
13	1

Fig. 1. Example of a small network: The Intercity railway (plus flyway) network in Germany approximatively shows a scale-free link distribution (see Figs. 2 and 3).

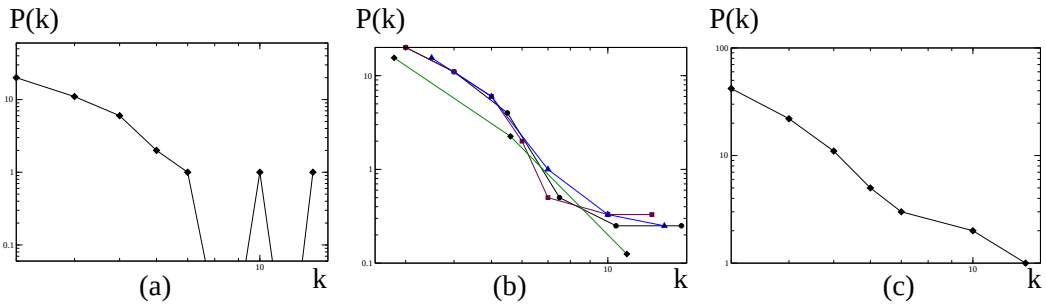


Fig. 2. (a) Problem of zeroes (see text). (b) Result of different binnings depending on parameters c and $n_{\max}$. (c) The Integrated density is defined free of parameters.

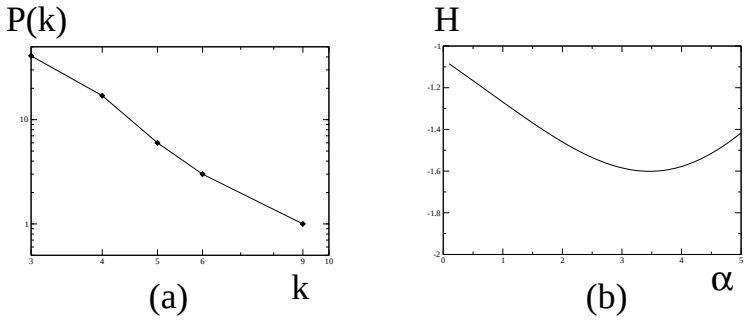
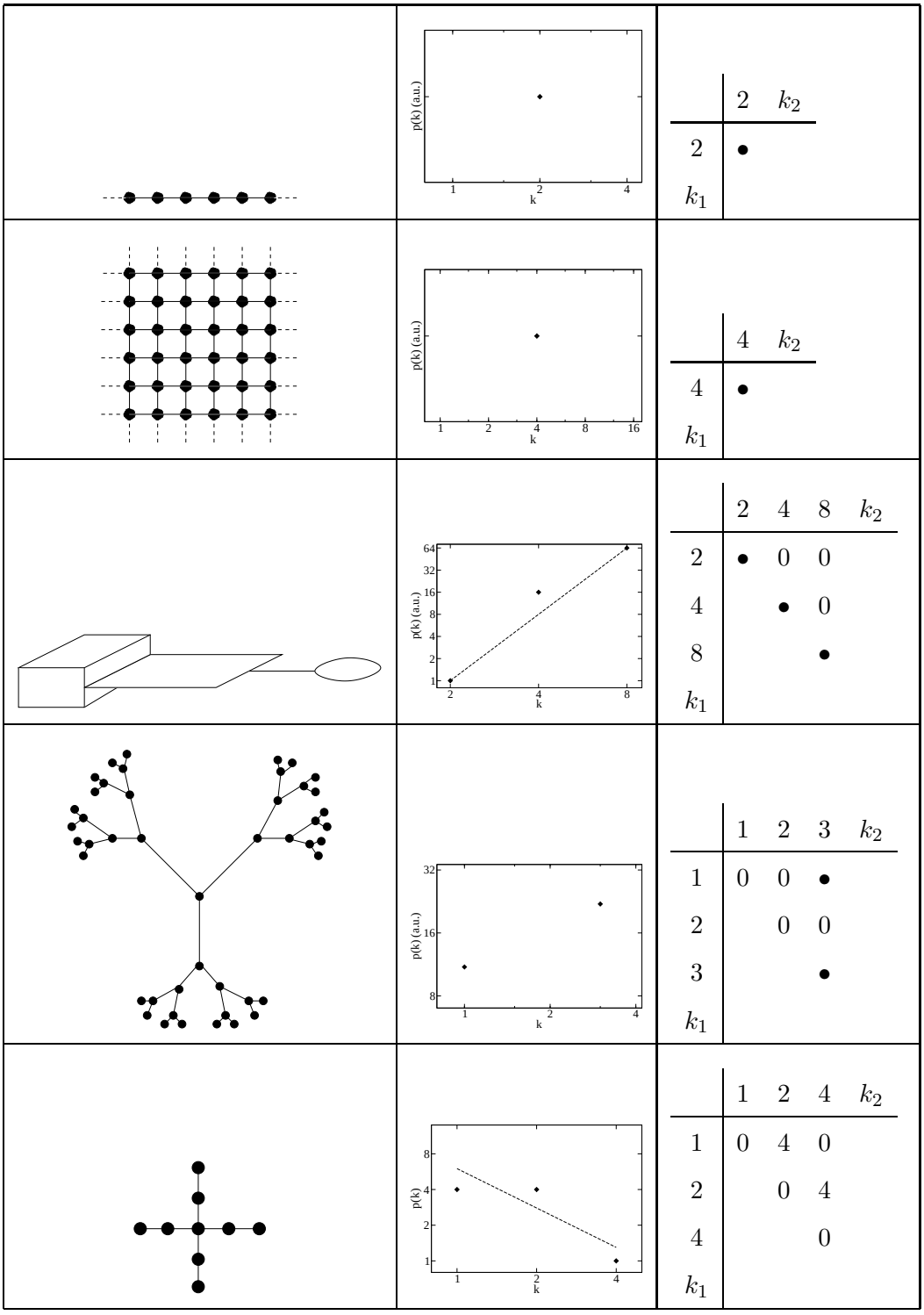


Fig. 3. (a) The biased entropy of the distribution shows an extremum with respect to the exponent α (b). From here, we have a parameter-free estimation of α.



	2	k_2
2	•	
k_1		

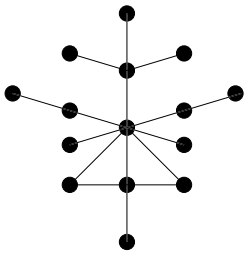
	4	k_2
4	•	
k_1		

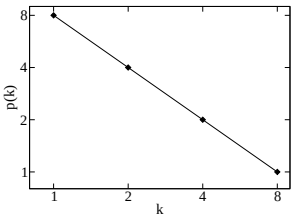
	2	4	8	k_2
2	•	0	0	
4		•	0	
8			•	
k_1				

	1	2	3	k_2
1	0	0	•	
2		0	0	
3			•	
k_1				

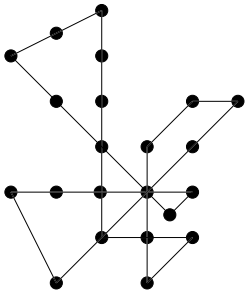
	1	2	4	k_2
1	0	4	0	
2		0	4	
4			0	
k_1				

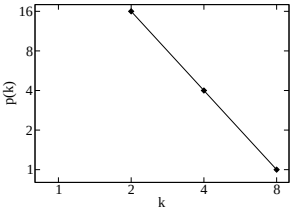
Fig. 4. Small non-complex networks: These networks are large, not complex, and not scale-free. A single entry or a single diagonal with nonzero entries indicates low complexity. Shown are a regular lattice in 1D and 2D (top) and a Bethe lattice and a star graph (bottom) The third example (middle) is the box-plane-stick-loop concatenation of different-dimensional finite lattices, widely used as data analysis test set.



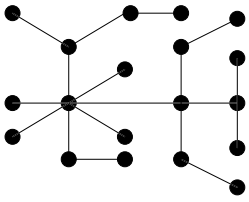


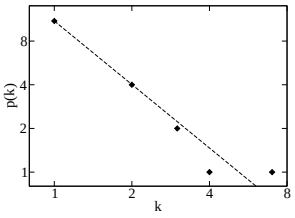
	1	2	4	8	k_2
1	0	2	4	2	
2		0	2	4	
4			0	2	
8				0	
k_1					



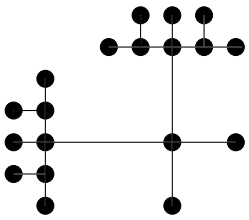


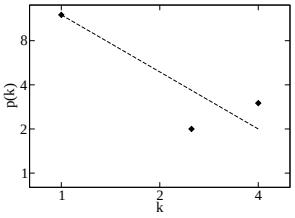
	2	4	8	k_2
2	11	6	4	
4		3	4	
8			0	
k_1				





	1	2	3	4	5	6	7	k_2
1	0	4	3	0	0	0	4	
2		0	1	2	0	0	1	
3			0	1	0	0	1	
4				0	0	0	1	
					0	0	0	
						0	0	
k_1							0	





	1	2	3	4	k_2
1	0	8	0	4	
2		0	0	4	
3			0	0	
4				2	
k_1					

Fig. 5. Small complex networks: A striking observation is that entries are quite evenly spread on the offdiagonals. Can this be used to define a complexity measure?

4 Definition of the Offdiagonal Complexity (OdC)

Let g_{ij} be the adjacency matrix of a graph with N nodes, i.e., $g_{ij} = 1$ if nodes i and j are connected, else $g_{ij} = 0$. Then OdC is defined as follows [15].
(i) For each node i, let $l(i)$ be the node degree, i.e. the number of edges (links),

$$l(i) := \sum_{j=0}^{N-1} g_{ij} \tag{2}$$

(ii) Let c_{mn} be the number of edges between al pairs of nodes i and j, with node degrees $m = l(i)$, $n = l(j)$ with $l(j) \geq l(i)$ (ordered pairs), i.e.,

$$c_{mn} := \sum_{j=0}^{N-1} \sum_{j=0}^{N-1} g_{ij} \delta_{m,l(i)} \delta_{n,l(j)} H(l(i) - l(j)). \tag{3}$$

Here δ is the Kronecker symbol and $H(x) = 1$ for $x \leq 0$ and $H(x) = 0$ for $x < 0$. Due to the pair odering, the matrix c_{mn} has entries only on the main diagonal and above. Thus, c_{mn} is a (not normalized) node-node link correlation matrix.
(iii) Summation over the minor diagonals, or offdiagonals, i. e. all pairs with same $k_i - k_j$ up to $k_{\max} = \min_i\{l(i)\}$, and normalization,

$$\tilde{a}_k = \sum_{i=0}^{k_{\max}-k} c_{i,k+i}, \qquad A := \sum_{k0}^{k_{\max}} \tilde{a}_k, \qquad \forall_k a_k := \tilde{a}_k / A. \tag{4}$$

(iv) Then OdC is defined as an entropy measure on this normalized distributions (here it is understood that $0 \ln(0) = 0$),

$$\mathrm{OdC} = -\sum_{k=0}^{k_{\max}} a_k \ln a_k. \tag{5}$$

OdC is an approximative complexity estimator that takes as values zero for a regular lattice (an orthogonal n-dimensional lattice with periodic boundaries consists of bulk nodes with $2n$ neighbors. Thus $c_{2n,2n} = 1$ is the only entry; for this regular structure OdC vanishes. Also a 2-dimensional hexagonal lattice has only one entry), zero for a fully connected graph, low values for a random graph, and higher values for 'apparently complex' structures. One main advantage is that it does not involve costly (high-order or NP-complete) computations.

5 Application to the Helicobacter pylori protein interaction graph and reshuffling to a random graph

To demonstrate that OdC can distinguish between random graphs and complex networks, the Helicobacter pylori protein interaction graph [20] has been chosen. For different rewiring probabilities p and 10^2 realizations each, the links have been reshuffled, ending up with a random graph for $p = 1$. As can be seen in Fig. 6, rewiring in any case lowers the Offdiagonal Complexity.

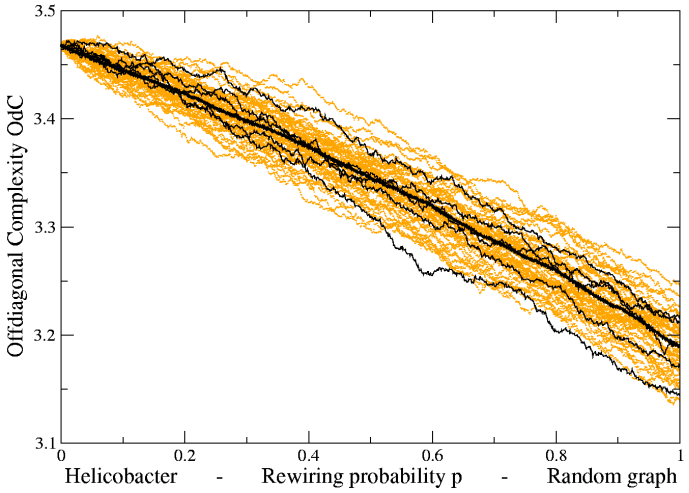


Fig. 6. OdC for random reshufflings of the *Helicobacter pylori* network (left, $p = 0$) up to a rewiring probability of $p = 1$ (right). The bold line shows the average, five OdC trajectories along a rewiring path are shown for illustration (thin lines).

6 Conclusions and Outlook

A new complexity measure for graphs and networks has been proposed. The motivation of its definition is twofold: One observation is that the binning of link distributions is problematic for small networks. Herefrom the second observation is that if one uses instead of the (plain) entropy of link distribution, which is unsignificant for scale-free networks, a "biased link entropy", it has an extremum where the exponent of the power law is met.
The central idea of OdC is to apply an entropy measure to the degree correlation matrix, after summation over the offdiagonals. This allows for a quantitative, yet still approximative, measure of complexity. OdC roughly is 'hierarchy sensitive' and has the main advantage of being computationally not costly.

Acknowledgments. J.C.C. thanks Christian Starzynski for providing the simulation code for Fig. 6, and an anonymous referee for constructive remarks.

References

[1] D.J. Watts and S.H. Strogatz, Nature 393, 440-442 (1998).

[2] A.L.Barabasi and R. Albert, Science, 286, 509-512 (1999).

[3] R. Albert, A.L.Barabasi, Statistical mechanics of complex networks, Rev.Mod.Phys., 74,47. (2001).

[4] M.E.J. Newman, The structure and function of complex networks, SIAM Review 45, 167 (2003).

[5] S.N. Dorogovtsev, J.F.F. Mendes, Evolution of networks, Adv. Phys. 51, 1079 (2002).

[6] S. Bornholdt, H.-G. Schuster (eds.), Handbook of Graphs and Networks, Wiley-VCH, Berlin (2002).

[7] H. A. Ceccatto and B. A. Huberman, Physica Scripta 37, 145 (1988).

[8] M. Gell-Mann, S. Lloyd. Information measures, effective complexity, and total information. Complexity 2(1), 44-52 (1996).

[9] M. Gell-Mann. What is complexity? Complexity 1(1), 16-19 (1995).

[10] T. Wilhelm, An elementary dynamic model for nonbinary food-webs, Ecol.Model. 168, 145. (2004).

[11] H. Meyer-Ortmanns, Functional Complexity Measure for Networks, Physica A 337, 679 (2004).

[12] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, Science 298, 824 (2002).

[13] B. Machta and J. Machta, Parallel dynamics and computational complexity of network growth models, Phys. Rev. E, 71, 026704 (2005).

[14] Jens Christian Claussen, AKSOE 3.10, Verhandl. Deutsche Phys. Ges., Regensburg (2004) (Extended version of unpublished talk draft, Nov. 11, 2003).

[15] Jens Christian Claussen, Proc. ECMTB05, Dresden, in print.

[16] Maurice George Kendall and Jean Dickson Gibbons, Rank Correlation Methods, 5^{th} ed. Edward Arnold, London (1990).

[17] Railway network of Deutsche Bahn AG, `http://www.bahn.de`

[18] P. Grassberger, Entropy Estimates from Insufficient Samplings, arXiv.org/abs/physics/0307138 (2003).

[19] D.H. Johnson and S. Sinanovic, Symmetrizing the Kullback-Leibler distance, unpublished (2001).

[20] Helicobacter pylori data, `http://www.cosin.org/`, `http://www.helico.com/`