

Chapter 1

On Motif Statistics in Protein Spatial Networks

Blake Stacey

In recent years, networks derived from complex systems have been studied not just in terms of global properties like degree distributions, but also in terms of motifs, small subgraphs whose appearances can be examined statistically. Motifs which occur more often than chance predicts are often presumed to indicate some feature of local structure which is preferred for biological, physical or geometrical reasons. We test a claim made in R. Milo et al., *Science* 303 (5 March 2004) to the effect that protein structures can be approached in this way, studying not three proteins but a set of 830. Overall, the general claim of the earlier paper is borne out: the spatial distribution of secondary-structure elements can be roughly understood as a geometrically constrained network. Structures of individual proteins are reflected in the clustering coefficients of the networks derived from the protein geometries. The investigation of a network growth model based on geometric constraints reveals a conceptual link with entropy-estimation techniques invented for high-energy physics.

1.1 Introduction

Connecting points with lines has been a way of thought for a very long time. In the fourteenth century, theologian Ramon Llull tried to understand Jehovah with a complete 8-graph whose nodes were labeled *Veritas*, *Gloria* and so forth[4]. In 1978, science historian James Burke described our technological society as a network, “each part of which is interdependent with all the others,” so that a failure in one place—say, an electrical relay in the Adam Beck Two power station at Niagara Falls—can have severe effects far away, like the Great Northeastern Blackout (compare [1, 3, 16]). Burke used the network as a metaphor, a tool for exploring how our society sustains itself as well as the insolubly wedded question

of how our science and technology advance[5]. In later years, quantitative data became available on a wide variety of complex worldly affairs: protein interactions, organism genomes, the global computer web Burke described in its earliest stages, and more. The presence of this interdisciplinary data has made possible interdisciplinary research, and the apparent detection of ubiquitous features has been greeted with wide acclaim. After the discovery of the scale-free hammer, we have been able to treat all observable systems as nails.

Much discussion has transpired about *global* properties of networks which live in abstract spaces, but it is also interesting to look at *local* properties of graphs embedded in familiar, Euclidean space. We may expect that network models will be most useful in studying those systems whose fundamental parts are all alike or described by few parameters apiece, say a single numerical weight per node, with all other information encoded in their interrelationships. We treat all cities as interchangeable, or labeled only by their population. Whether this is a reasonable approximation can only be determined by assaying its predictive power and its correspondence with experimental observations. (Likewise, we often neglect the distinguishing features of the connections, drawing them all in the same color and with the same thickness.) Here, proteins were analyzed at two levels. First, individual amino-acid residues were treated as nodes and connected if their C_α atoms were closer than 8 Å (per [11]). Then, in the second approach, secondary-structure elements were mapped onto vertices. Each α -helix or β -sheet became a node, and nodes were connected if their corresponding structure elements came within 10 Å of each other. α -helices and β -sheets are three-dimensional objects, of course, so the distance between them is defined (with appropriate arbitrariness) to be the minimum separation between C_α atoms of their respective amino acids. This information can be computed from the protein's PDB file with relative ease.

One method to study these local properties is to hunt through it for repeated *motifs*, subgraphs containing a few nodes apiece, in the hope that a subgraph which occurs many times is physically or biologically significant[14]. Milo *et al.* advance the claim that a particular set of four-node motifs are “overrepresented” in protein structure; that is, these particular motifs occur in protein geometry more often than expected by chance. However, this claim was only supported by three proteins[13]. In this paper, we shall examine how well it applies to 830 more.

1.2 Sequence Profiles

Given a network (produced by whatever means), we can count the number of times a particular subgraph occurs. Take a set of motifs labeled by i , and let the number of occurrences for each motif be n_i . (In this paper, $i = 1, \dots, 6$.) The quantities n_i are not the most convenient way to measure the abundances of network motifs, since they scale with the overall network size. Besides, the quantity of real interest is not how many times a particular grouping of nodes occurs, but how many occurrences we find *compared to what we would observe due*

to random chance. Imagine a randomized ensemble of networks, each with the same degree profile as the original, but with the edges rearranged. Presumably, if a particular motif is prevalent (or lacking) for biological reasons, it will occur more (or less) frequently than in this randomized ensemble. (In a soap opera of order N , how many love triangles do we expect to see?) We define a measure of the statistical significance of a particular n_i , relative to chance expectation:

$$Z_i \equiv \frac{n_i - \langle r_i \rangle}{\sqrt{\langle r_i^2 \rangle_c}}. \quad (1.1)$$

Here, $\langle r_i \rangle$ denotes the mean number of times motif i occurs per network, averaged over all networks in the ensemble. The cumulant in the denominator, $\langle r_i^2 \rangle_c$, is the variance of r_i , likewise computed over all the randomized networks.

The networks in the randomized ensemble were generated by swapping edges in the network generated from the protein geometry. This method has the advantage that it preserves connectivity properties of the network: each node will have the same number of ingoing and outgoing connections as it did in the original.

To compare Z -scores calculated for different proteins, we introduce a new variable, S_i , which is normalized so that the sum over all i is unity:

$$S_i \equiv \frac{Z_i}{\sqrt{\sum_i Z_i^2}}. \quad (1.2)$$

The set $\{S_i\}$ is termed the *sequence profile*. Alternatively, it may be preferable to use a different measure of statistical significance, which we shall term the D -score:

$$D_i \equiv \frac{n_i - \langle r_i \rangle}{n_i + \langle r_i \rangle + \epsilon}. \quad (1.3)$$

Here, ϵ is a constant of order unity chosen at the investigator's convenience. (To compare our results directly with those of Milo *et al.*, we have made our calculations with $\epsilon = 4$.) D_i can naturally be normalized in the same way as Z_i .

This method suffers from several drawbacks. Consider, for example, an arrangement such as the network of neurons in an animal brain, where it is difficult for nodes separated by a large physical distance to connect. In this case, we might find groups of neighboring nodes connected more strongly than chance expectations only because they are situated near each other, and not for any reason having to do with biological selection. A "toy model" has been constructed, in which nodes are stochastically connected to proximate neighbors (connections being formed with a probability that falls off as a Gaussian). Analyzing the resulting networks shows the same prevalent motifs as Milo *et al.* find in the *C. elegans* neural connection structure[18]. Such toy models do, however, show motifs which do *not* appear in "real" networks, indicating that the sequence-profile method is not entirely fragile, as long as the *entire* sequence profile is studied[15].

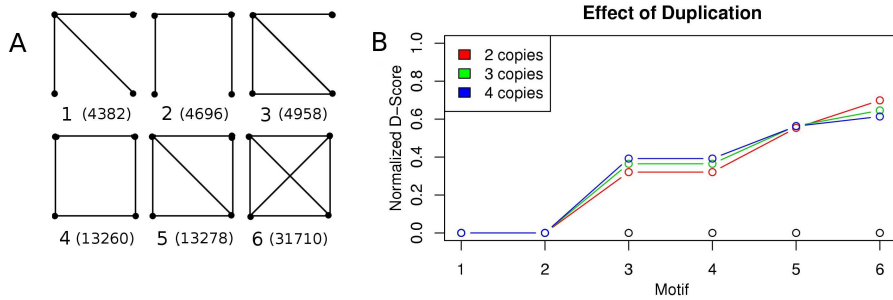


Figure 1.1: (A) The six motifs which Milo *et al.* explore in protein structure, with their MFinder[17] ID codes. Motifs 3, 5 and 6 are claimed to be “overrepresented”. (B) Spurious sequence profile produced by setting $n_i = \langle r_i \rangle$ and then scaling.

Another issue arises when the network under scrutiny exhibits *symmetry*. Consider a protein which is made of multiple peptide chains, organized into *domains* such that the protein is composed of perhaps four closely similar pieces. Neglecting the “edge effects” of one domain’s α -helices touching another’s, the network for the entire protein is just the graph of the domain, duplicated without overlaps.¹ Duplicating by a factor d amplifies each n_i by d , but it also affects the ensemble of randomized graphs we use for comparison. Because the degree profile is unaltered, the probability that a given subgraph plucked from a randomized graph will exhibit motif i is unchanged, but the number of ways to pick a subgraph grows with the network size. Using the results of [9], one can show that $\langle r_i \rangle$ scales as d^{n-l_i} , where n is the number of nodes in the motif (here 4) and l_i is the number of internal links. The net effect is to bias the sequence profile towards those motifs with higher l_i . If D_i is identically zero, the combination of symmetry and scaling creates a wholly spurious sequence profile.

1.3 Spatial Graphs

As indicated above, one complication is the issue of *geometry*. “Classic” scale-free (SF) graphs like the actor-film network are not embedded in space; what are the effects upon the motif profile of requiring that our graph be grown in a Euclidean world? Following the example of Herrmann, Barthélemy and Provero[6], we consider vertices scattered randomly throughout a unit volume according to a probability distribution $\rho(x)$. Vertices will be linked if they fall within a chosen distance R from one another. Graphs of this type occur in a variety of circumstances, from telecommunications to gene-expression microarrays, listed conveniently in [6]. In a more sophisticated model, vertices can be connected with a *probability* which depends upon their distance, such as a Gaussian fall-off.

¹The three proteins studied in [13] are all multi-domain polypeptides: an oxoreductase (PDB code 1AOR), a serine protease inhibitor (1EAW) and an immunoglobulin (1A4J).

The hard-sphere case which [6] studies explicitly can be taken as a reasonable first approximation.

Given a network presumably grown by a geometrical process, we would like to infer some characteristics of ρ from the statistical properties of the network. One such summary characteristic of ρ is the probability that two randomly chosen nodes will share an edge, knowing nothing else. If $\chi(x, y)$ is the “cutoff” function defining the proximity condition,

$$p = \int dx dy \rho(x) \chi(x, y) \rho(y). \quad (1.4)$$

Another property of interest is the *clustering coefficient*, defined to be the probability that two neighbors of a randomly chosen node will themselves be directly linked. In the spatial model, C is given by the probability that two points x_2 and x_3 randomly chosen within the “acceptance zone” of x_1 are themselves close enough to be linked:

$$C = \int dx_1 dx_2 dx_3 \rho(x_1) \chi(x_1, x_2) \rho(x_2) \chi(x_1, x_3) \rho(x_3) \chi(x_2, x_3). \quad (1.5)$$

If we know p and C , we can estimate the probability that a small subgraph of n nodes drawn from the full network of N will exhibit a particular motif.² We ask with what likelihood each connection in the motif will be drawn, and we multiply these factors together; for a 4-node motif, a total of six links could possibly be drawn, and thus the total probability will be some combination of six factors taken from the set $\{p, (1-p), C, (1-C)\}$. The specific combination is determined by the motif topology. We can estimate the quantity of each motif seen by multiplying the expressions in the table by $\binom{N}{n}$.

As the authors of [6] note, it is possible to grow a scale-free (power-law) network embedded in Euclidean space by a judicious choice of $\rho(x)$. The assumption that p and C are roughly sufficient to characterize ρ is equivalent to saying that ρ is “well-behaved” in such a way that this will not happen; further discussion of this point lives in [8] and the Appendix.

We constructed a Python program to implement this model of graph formation. The program randomly chooses points within a 3D unit volume, makes pairwise connections based on a user-defined threshold, calculates summary statistics and outputs the network in **MFinder**-ready format[17]. Fig. 1.2(C) shows the results, the salient feature being that the model predictions correspond with the motifs as measured by **MFinder**.

1.4 Protein Analysis

The dataset was obtained from Uwe Hobohm and Chris Sander’s PDB_SELECT list of Protein Data Bank chain identifiers.[7] PDB chain IDs in PDB_SELECT are

²This is done “between the lines” in [13]; see also [8].

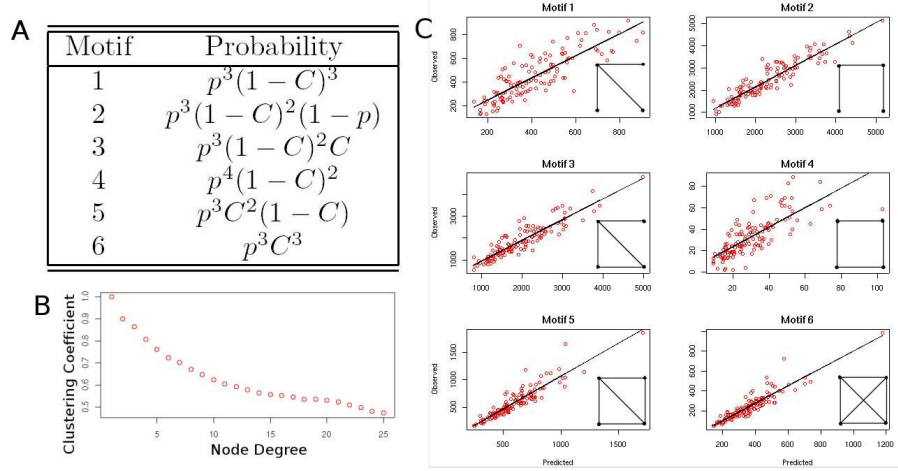


Figure 1.2: (A) Table of motif occurrence probabilities. (B) Average clustering coefficient per node as a function of node degree for randomly-generated spatial graphs. (C) Motif counts for randomly generated graphs, as measured by **MFinder** and predicted theoretically.

chosen such that no two peptide sequences are more than a chosen percentage homologous; sequences used here were taken from the 25% list. Only one domain from each protein was used. For AA-level analysis, the largest α -helix and the largest β -sheet were drawn from each protein; in the second case, all secondary-structure elements in the domain marked in **PDB.SELECT** were converted to vertices. A Python program was used to extract the pertinent information from the PDB files, construct the corresponding graphs and output them in a format that the **MFinder** code is able to understand. The Python code accepts the threshold distance (in Å) as a user-defined parameter, and the analysis was conducted at thresholds of 10, 15 and 20 Å. Other Python scripts, Octave and R were used to break down the results.

Is the sequence-profile technique powerful enough to distinguish helices from sheets? Fig. 1.6(A) below indicates that it is; the distinct clouds of points for all but the most heavily connected motif strongly hint that the two “families” of structure can be separated by their motif statistics.

Several attributes of the secondary-structure networks are worth discussion. The first datum to note is that the degree profiles (averaged over all graphs) calculated at each of the three thresholds are roughly Poissonian in comportment, as shown in Fig. 1.3. This correlates well with the results of Li *et al.*[11], in which the authors study a set of 424 protein chains at the individual amino-acid level, treating each residue as a node and linking nodes whose C_α atoms are closer than 8 Å.

Second, as was also seen in [11], the clustering coefficients for the protein-structure networks are much larger than those for Erdős-Rényi (ER) graphs of

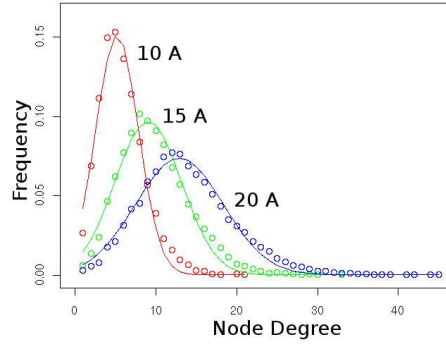


Figure 1.3: Degree profiles fit with Gaussians.

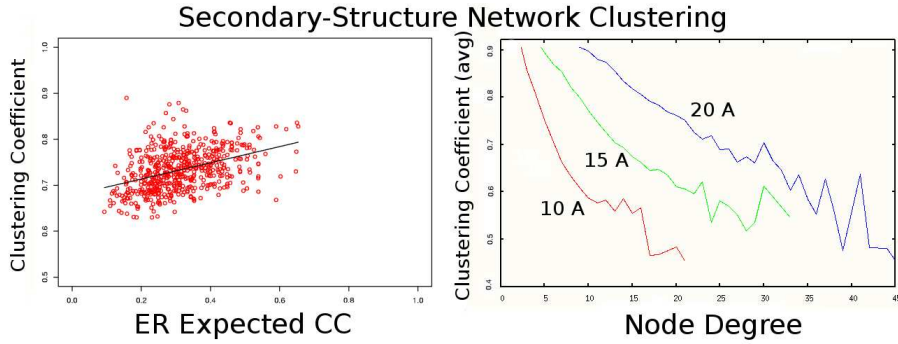


Figure 1.4: On the left, clustering coefficients for the protein-structure networks compared to those for ER graphs of the same size and average connectivity. The regression line is $y = 0.68(\pm 0.01) + 0.18(\pm 0.02)x$, with $r = 0.4$. On the right, $C(k)$ for individual vertices plotted as a function of vertex degree, averaged over all proteins.

the same size and average degree. Looking more closely, we find that although C for the real proteins is greater than the ER prediction, the two are not unrelated: knowing only the ER value of C (that is, N and $\langle k \rangle$), one can linearly approximate the true value. Fig. 1.4 displays the correlation. This was not studied in [11]’s residue-level analysis. Third, we find that the clustering coefficient for vertices having the same k , when averaged over all proteins, falls off with k . SF researchers typically see a decay $C \propto k^{-1}$ as evidence that the graph is hierarchical. Note that increasing the threshold from 10 Å turns the decay more and more linear, obscuring as one might expect the signatures of organization.

Next, we turn to the motif-statistical properties of these protein structures. Fig. 1.5 shows the histograms for the normalized D -scores of the six subgraphs under investigation. The general pattern seen in [13]’s three proteins is borne out: motifs 3, 5 and 6 appear more often in the real proteins than they do in the randomized ensembles. The blue dots indicate the average results from a set of

100 randomly-generated graphs with p and C typical for the protein networks; their agreement is generally better than the analytic p -and- C model. Note that the plots in Fig. 1.5 show this particular sequence profile becoming more clearly evident as N increases.

We ask, naturally, if the networks' sequence profiles show any variation "on top of" the values we expect on geometrical grounds. In Fig. 1.6(B), we plot the normalized D -scores for each motif against that calculated with the model in §1.3. In each case, the observed D_i are positively correlated with the spatial-graph expectations, with predictive ability comparable to the generated networks discussed above. The geometrical model is, overall, good at explaining the motif abundances in protein networks.

1.5 Conclusions and Acknowledgments

The motif-statistic properties of 833 protein domains, drawn from the PDB_SELECT representative list, were investigated and found to follow the expectations for random graphs grown in space. The motifs found to be overabundant with respect to chance (that is, with positive D_i) are the same ones identified in [13]; in fact, though they do not give quantitative specifics, the authors of [11] report seeing the same sequence profile in the protein-structure networks they made at the residue level. This provides a pleasing consistency and suggests that the effects of "coarse-graining" merit further theoretical exploration. Random spatial graphs are a reasonable starting point for later graph-theoretic studies of protein structure.

I would like to thank Leonid Mirny of the Harvard-MIT Divison of Health Sciences and Technology for introducing this problem to me. Eric Downes provided suggestions during the investigation and the article preparation; at ICCS 2006, Franziska Matthäus suggested the analysis which became Fig. 1.6(A).

1.6 Appendix: Spatial Scale-Free Graphs and Entropy

Suppose we have a node placed at x . With what probability will it have a given number k of neighbors? This is just the probability to find k nodes within the volume near x . Introduce a cutoff function $\chi(x, y)$ which specifies the region in which nodes are sufficiently close to be connected; the distance dependence of χ may be a step function, indicating a hard-sphere cutoff, or it might be a softer relation like a Gaussian. The probability that a second node is placed within the acceptance region of a node at x is

$$q(x) = \int dy \chi(x, y) \rho(y). \quad (1.6)$$

(This reduces to Eq. (4) of [6] in the hard-sphere case.) Analytic results are easiest to derive in the limit that $N \rightarrow \infty$. To keep the results of interest

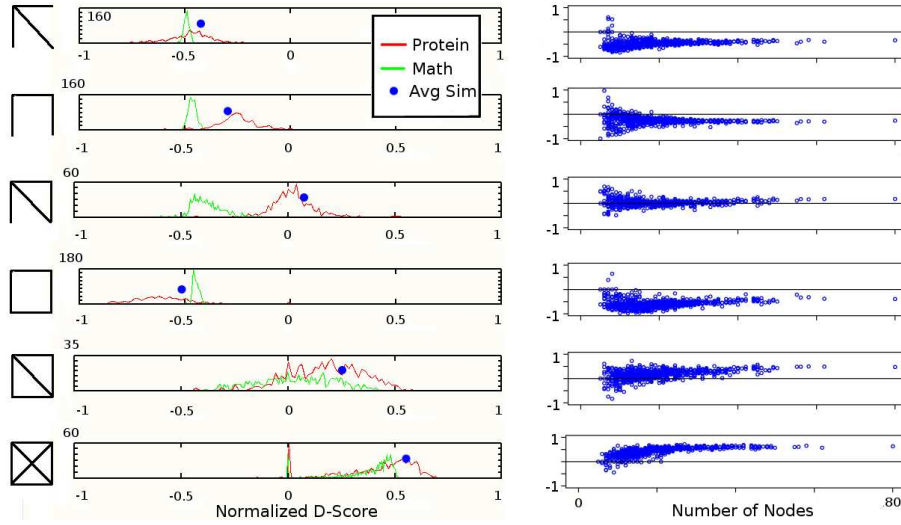


Figure 1.5: On the left, summary histogram of D -scores for the six 4-motifs under study. On the right, D -scores plotted against the number of nodes for the protein networks (computed at a 10-Å threshold).

finite, we must simultaneously take the limit $R \rightarrow 0$ (or shrink the corresponding scale parameter in χ , whatever it may be). We wish to have the product $N \int dx dy \chi(x, y)$ tend to a finite constant, which following the earlier paper we term α . In this event, the expected number of nodes found within the region defined by χ remains finite and tends to $\alpha \rho(x)$.

When we take the “thermodynamic limit”, we can simplify our integrals and average over space to obtain the result in [6],

$$P(k; \alpha) = \frac{\alpha^k}{k!} \int dx \rho^{k+1}(x) e^{-\alpha \rho(x)}. \quad (1.7)$$

In the “thermodynamic limit”, χ tends to a delta function and the clustering coefficient Eq. (1.5) becomes

$$C = \alpha^3 \int dx \rho^3(x). \quad (1.8)$$

The authors of [6] demonstrate that a random spatial graph produced from a $\rho(x)$ satisfying certain properties can be “scale-free”. Here, “cumulative advantage” or “preferential attachment” is only an artifact of location: nodes near a peak of $\rho(x)$ will tend to be more highly connected since they are likely to have more neighbors, but this same region is also the most likely place for new vertices to appear. (The rich get richer because of where they live, not who they know!) They define an SF graph to be one in which the moments $\langle k^\nu \rangle$ of the

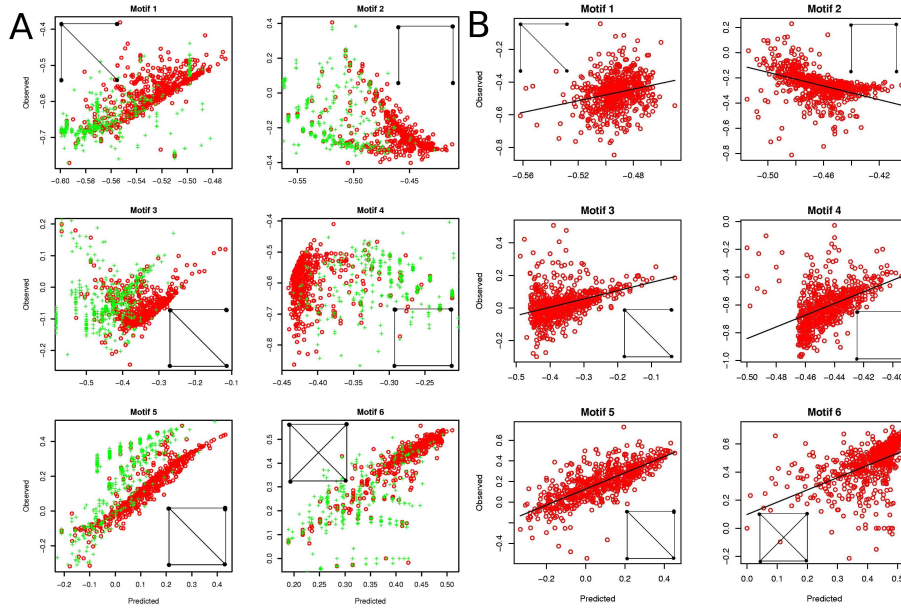


Figure 1.6: D -scores for protein structure networks compared to the predictions of the random spatial graph model, using p and C for each graph to compute an expected $\{D_i\}$. (A) Networks computed at AA-residue level; α -helices in red, β -sheets in green. (B) Networks computed at secondary-structure resolution. Values of the correlation coefficient $r = 0.24, 0.43, 0.38, 0.46, 0.78$ and 0.55 .

degree distribution diverge for all ν greater than some ν_{\max} ; it is easy to see how this relates to a power-law decay at large k . Applying textbook knowledge of combinatorics to Eq. (1.7), they derive the following expression for $\langle k^\nu \rangle$ in terms of α :

$$\langle k^\nu \rangle_\alpha = \sum_{m=0}^{\nu} \mathbb{S}_\nu^{(m)} \alpha^m \int dx \rho^{m+1}(x). \quad (1.9)$$

Here, $\mathbb{S}_\nu^{(m)}$ denotes the *Stirling numbers of the second kind*, which count the number of ways to partition a pile of ν elements into m non-empty boxes. The integral in Eq. (1.9) measures the information content of ρ and is closely related to the Rényi entropy[10], by

$$R_q = \frac{1}{1-q} \log \int dx \rho^q(x). \quad (1.10)$$

If one knows the Rényi entropies for values of $q > 1$, one can find the Shannon entropy by extrapolating the (q, R_q) curve down to $q = 1$. According to the definition in [6], then, a graph is SF if the Rényi entropies diverge for all $q > \nu_{\max} + 1$.

An interesting parallel exists between this model of random spatial graphs

and a method developed to estimate the entropy of high-energy particle collisions. This technique originated with the work of Shang-Keng Ma[12], who considered the problem of how one could compute an entropy during a Monte Carlo simulation. For the moment, consider a system prepared at a fixed energy \mathcal{E} . The entropy of this system is given by the phase-space volume through which the system's trajectory passes. Ma proposed sampling the system at a number of points during its evolution and counting the "coincidences", that is, the number of times two sample points fall within the same bin. Ma indicates that, for the fixed-energy case, the entropy is given by the simple relation

$$S = k_B \log(\text{volume}) = -k_B \log C_2 \quad (\text{fixed } \mathcal{E}). \quad (1.11)$$

Bialas and Czyz[2] provide a generalization to the "canonical" case of non-fixed energy; their phenomena of interest are multi-particle interactions at high energies, including dense hadronic matter and quark-gluon plasmas. The fundamental notion is to correct Eq. (1.11) by including higher-order coincidences, three or more configurations landing within the same sample bin.

In this approach, one counts the number of sample points included within small phase-space regions, while in studying spatial graphs, one connects nodes which are separated by small distances. A "coincidence" is not dissimilar to a connected subgraph. The parallel is clear enough that it is not surprising Bialas and Czyz derive a formula for the overall entropy in terms of the Rényi entropies.³

We can find the probability of a ν -fold coincidence from the distribution function ρ : calculate the probability of observing ν vertices in a volume \mathbb{V} and then integrate over all choices of \mathbb{V} . Using the cutoff function χ introduced earlier,

$$C_\nu = \int dy \left[\int dx \chi(x, y) \rho(x) \right]^\nu. \quad (1.12)$$

In the "thermodynamic" limit, χ tends to a delta function, and C_ν becomes

$$C_\nu = \alpha^\nu \int dy \rho^\nu(y) = \alpha^\nu \langle \rho^{\nu-1} \rangle. \quad (1.13)$$

Note that C_3 is just the clustering coefficient of Eq. (1.8).

Bibliography

- [1] ALBERT, R., H. JEONG, and A.-L. BARABÁSI, "Attack and error tolerance of complex networks", *Nature* **406** (2000), 378–382.
- [2] BIALAS, A., and W. CZYZ, "Event by event analysis and entropy of multiparticle systems", *Phys. Rev. D* **61** (2000).

³They attribute the argument to K. Zyczkowski.

- [3] BOLLOBÁS, B., and O. RIORDAN, “Robustness and vulnerability of scale-free random graphs”, *Internet Math.* **1** (2003), 1–35.
- [4] BORGES, Jorge Luis, *Ramon Llull’s Thinking Machine*, Viking, (1999), pp. 155–59.
- [5] BURKE, J., “Connections” (1978), See also *Connections* (Macmillan, 1978), ISBN 0-333-24827-9.
- [6] HERRMANN, C., M. BARTHÉLEMY, and P. PROVERO, “Connectivity distribution of spatial networks”, *Phys. Rev. E* **68**, 026128 (2003).
- [7] HOB OHM, U., M. SCHARF, R. SCHNEIDER, and C. SANDER, “Selection of a representative set of structures from the brookhaven protein data bank”, *Protein Science* **1** (1992), 409–417, July 2005 update from <http://bioinfo.tg.fh-geissen.de/pdbselect>.
- [8] ITZKOVITZ, S., and U. ALON, “Subgraphs and motifs in geometric networks”, *Phys. Rev. E* **71**, 026117 (2005).
- [9] ITZKOVITZ, S., R. MILO, N. KASHTAN, G. ZIV, and U. ALON, “Subgraphs in random networks”, *Phys. Rev. E* **68**, 026127 (2003).
- [10] JIZBA, P., and T. ARIMITSU, “Observability of Rényi’s entropy”, *Phys. Rev. E* **69**, 2 (2004).
- [11] LI, J.-J., D.-S. HUANG, T.-M. LOK, M. R. LYU, Y.-X. LI, and Y.-P. ZHU, “Network analysis of the protein chain tertiary structures of heterocomplexes”, *Protein and Peptide Letters* **13** (2006), 407–412.
- [12] MA, Shang-Keng, *Statistical Mechanics*, World Scientific (1985).
- [13] MILO, R., S. ITZKOVITZ, N. KASHTAN, R. LEVITT, S. SHEN-ORR, I. AYZENSHTAT, M. SHEFFER, and U. ALON, “Superfamilies of evolved and designed networks”, *Science* **303** (2004), 1538–1542.
- [14] MILO, R., S. SHEN-ORR, S. ITZKOVITZ, N. KASHTAN, D. CHKLOVSKII, and U. ALON, “Network motifs: Simple building blocks of complex networks”, *Science* **298** (2002), 824–827.
- [15] R. MILO, et al., “Response to comment ...”, *Science* **305** (2004), 1107d.
- [16] SHARGEL, B., H. SAYAMA, I. R. EPSTEIN, and Y. BAR-YAM, “Optimization of robustness and connectivity in complex networks”, *Phys. Rev. Lett.* **90**, 6 (2003), 068701.
- [17] U. ALON, et al., “MFinder program” (2002), <http://www.weizmann.ac.il/mcb/UriAlon/groupNetworkMotifSW.html>.
- [18] Y. ARTZY-RANDRUP, et al., “Comment on ‘network motifs: Simple building blocks of complex networks’ and ‘superfamilies of evolved and designed networks’”, *Science* **305** (2004), 1107c.