INSTITUTE OF PHYSICS PUBLISHING JOURNAL OF PHYSICS A: MATHEMATICAL AND GENERAL

J. Phys. A: Math. Gen. **38** (2005) 9599–9620 doi:10.1088/0305-4470/38/44/001

Analytical solution of a stochastic content-based network model

Muhittin Mungan[1,2], **Alkan Kabakçıoğlu**[3,4], **Duygu Balcan**[5] **and Ayşe Erzan**[2,5]

[1] Department of Physics, Faculty of Arts and Sciences, Boğaziçi University, 34342 Bebek Istanbul, Turkey
[2] Gürsey Institute, PO Box 6, Çengelköy, 34680 Istanbul, Turkey
[3] Department of Physics, Faculty of Arts and Sciences, Koç University, 34450 Sarıyer Istanbul, Turkey
[4] Dipartimento di Fisica, Università di Padova, I-35131 Padova, Italy
[5] Department of Physics, Faculty of Sciences and Letters, Istanbul Technical University, Maslak 34469, Istanbul, Turkey

Received 6 July 2005, in final form 23 September 2005
Published 19 October 2005
Online at stacks.iop.org/JPhysA/38/9599

Abstract

We define and completely solve a content-based directed network whose nodes consist of random words and an adjacency rule involving perfect or approximate matches for an alphabet with an arbitrary number of letters. The analytic expression for the out-degree distribution shows a crossover from a leading power law behaviour to a log-periodic regime bounded by a different power law decay. The leading exponents in the two regions have a weak dependence on the mean word length, and an even weaker dependence on the alphabet size. The in-degree distribution, on the other hand, is much narrower and does not show any scaling behaviour.

PACS numbers: 2.10.Ox, 89.75.Da, 89.75.Hc

1. Introduction

In a previous paper, two of us (Balcan and Erzan) [1] introduced and numerically simulated a content-based network [2] with random binary strings associated with each node. The network arose by postulating a directed edge to exist between the nodes i and j if and only if the string, which can be regarded as a random word associated with the ith node, occurred at least once in the random word associated with the jth.

The initial motivation for this model [1] was based on RNA interference [3], a mechanism where short RNA strings bind complementary sequences of mRNA to suppress their translation into protein molecules. It has since been found that RNA interference is in fact an integral part of gene regulation. [4] On the other hand, transcriptional regulation of gene expression also relies on the recognition of regulatory sequences (RS) by transcription factors (TF) which are themselves proteins coded by specific genes [5, 6]. In this case the sequence matching is less direct, being mediated by the short amino acid sequences on the TF which bind the RS, and binding is also subject to steric constraints [6–11]. Sequence similarity as the basis for pathogen recognition in mathematical models of the biological immune system has a long standing history [12] as well. Thus, sequence similarity seems to play an ubiquitous role in complex networks.

This stochastic network was shown [1] to display a distinctly different topology than either the classical random networks of Erdös and Renyi [13] or the 'scale free' networks of the preferential-attachment universality class, introduced by Barabasi and Albert [14, 15]. Simulations [1] revealed that the in- and out-degree distributions, were markedly different, with in-degree distribution being rather localized. The out-degree distribution displayed a sharp crossover behaviour. For small out-degree d, the distribution $n(d)$ exhibited a putative scaling behaviour over a very narrow region, where the log–log plot could be fitted with a straight line with a slope $-\gamma_1 \simeq -1$, whereas, for larger d, log-periodic oscillations were found, with an envelope which could again be fitted, on a double logarithmic plot, by a linear graph with a slope $-\gamma_2 \simeq -1/2$.

The purpose of this paper is twofold. We first extend the model of Balcan and Erzan [1] to a broader class of models in which the random strings are derived from an $r+1$ letter alphabet and where partial matches are allowed. Second, we obtain analytical expressions for the ensemble averaged in- and out-degree distributions and investigate the crossover behaviour of the out-degree distribution. We show that the putative scaling behaviour observed in the simulations coincides with the leading power law behaviour obtained from our analytical results. We describe in detail the finite-size corrections to the infinite network limit. Comparison of our analytical predictions with the numerical data [1] for the $r = 2$ random bit string model with perfect matches shows very good agreement.

The paper is organized as follows: In the next section we reformulate the random string model of [1] for an alphabet of $r + 1$ letters. Our analytical results depend on the matching probability $p(l, k)$ that a string of length l selected randomly from the set of all strings of length l is contained at least once in a string of length k, $k \geqslant l$, that has been selected randomly from the set of all strings of length k. In section 3 we derive an approximate form for this probability that is valid for moderately long strings $k \lesssim r^l$ and that allows for partial matches. Using the results of section 3, we obtain in section 4 analytical expressions for the in- and out-degree distributions. We investigate the scaling behaviour of the out-degree distribution in these models and compare our results with the numerical data of [1]. We conclude this paper with a discussion of our results in section 5.

2. The random string model

Consider a random sequence C of fixed length L, consisting of letters from an alphabet A of $r + 1$ letters. The elements of the sequence $C, x \in \{0, 1, \ldots, r\}$ are assumed to be independently and identically distributed according to

$$P(x) = p\delta(x-r) + (1-p)\frac{1}{r}\sum_{m=0}^{r-1}\delta(x-m). \tag{1}$$

A subsequence G_i of C, composed of the letters $\{0, \ldots, r-1\}$ only, sandwiched between the ith and $(i + 1)$th occurrences of the letter 'r,' will be denoted as the ith 'random word,' or 'string,' and will be associated with the ith vertex of a graph. For convenience, we assume that a letter 'r' has also been placed at the 0th and the $(L+1)$th positions. With these definitions, the ith string can be written as

$$G_i = x_{i,1}, x_{i,2}, \ldots, x_{i,\ell_i}, \quad i = 1, 2, \ldots, N, \tag{2}$$

where N is the number of strings (equivalently, vertices), the 'letter' $x_{i,\lambda} \in \{0, r-1\}$, $\lambda = 1, \ldots, \ell_i$, and ℓ_i is the length of the ith string G_i. Let n_ℓ be the number of strings of length ℓ and $q = 1 - p$. It follows that

$$\sum_i \ell_i = L - N, \qquad \sum_\ell n_\ell = N, \tag{3}$$

$$\langle \ell \rangle = p^{-1} - 1, \qquad \langle n_\ell \rangle = Lp^2 q^\ell, \qquad \langle N \rangle = Lp. \tag{4}$$

Unless noted otherwise, we will assume that L and Lp are sufficiently large so that fluctuations in the number and length of the strings for different realizations of the random sequence C can be neglected when calculating statistical properties of quantities of interest. We will also discard the cases with $\ell = 0$ and construct the graph from the remaining vertices. The adjacency matrix is defined by the matching condition

$$w_{ij} = \begin{cases} 1 & G_i \subset G_j, \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

By $G_i \subset G_j$ we mean that there exists an integer λ such that $0 \leqslant \lambda \leqslant \ell_j - \ell_i$ and

$$x_{i,l} = x_{j,\lambda+l}, \qquad l = 1, \ldots, \ell_i. \tag{6}$$

Two vertices are said to be connected if the string G_i appears as a subsequence of G_j, or in other words G_j *matches* G_i. Thus $w_{ij} = 1$ indicates a directed link (an edge) from G_i to G_j. We will also consider *imperfect* matches, where equation (6) is valid only for some values of l rather than all values. In order to avoid ambiguity we will refer to the former case as a *perfect* match. For Lp large enough ($p > p_c(L)$, see [1]), which is assumed here, the graph consists of one giant cluster. We will henceforth refer to this graph as the network, and denote the vertices, or equivalently, the strings associated with them, as the 'nodes.'

The resulting network was numerically studied earlier by Balcan and Erzan in [1], for the case of binary strings, i.e., $r = 2$, and perfect matches equation (6), where it was shown that the logarithm of the out-degree distribution behaved linearly over a very narrow, initial range, with a slope of $\simeq -1$. Beyond a crossover point the distribution exhibited an oscillatory behaviour, whose envelope again behaved linearly on a log–log plot, with a different slope, namely $\simeq -1/2$. The out-degree distribution is shown in figure 1, where the numerical results were obtained [1] by averaging the out-degree distributions over 500 graphs, associated with independently generated sequences of length $L = 15\,000$, and $p = 0.05$. Note the strong oscillatory behaviour. It turns out that each peak in the out-degree distribution is supported predominantly by the out-degrees of genes with the corresponding common length l.

In order to proceed with the analytical treatment, it is convenient to group the G_i into subsets according to their lengths and we define

$$\mathcal{G}_l = \{G_i | \ell_i = l\}. \tag{7}$$

It turns out that the central quantity determining the behaviour of the in- and out-degree distributions is the probability $p(l, k)$ that a string in $\mathcal{G}_l$ has an outgoing edge terminating in a member of $\mathcal{G}_k$. We therefore turn next to the derivation of $p(l, k)$. The discussion of the degree distributions will then be taken up in section 4.

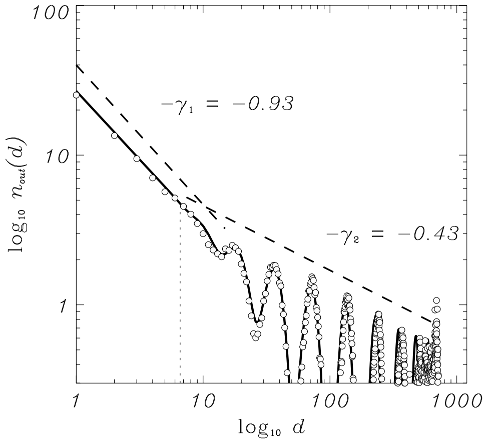


Figure 1. Scaling behaviour of the out-degree distribution. The numerical data (circles) show a crossover in the scaling behaviour from small values of the out-degree to larger values. The solid line is the theoretical expression. The dashed lines serve as a guide to the eye for the predicted scaling behaviour and have been offset for clarity. The crossover occurs at $d_c = 6.6$ and has been shown as a vertical line.

3. Analytical results for the matching probability

Let x and y be variables such that $x, y, \in \{0, \ldots, r-1\}$. Define an interaction $u(x, y)$ between x and y as

$$u(x, y) = 1 - \delta(x - y). \tag{8}$$

Let $\mathbf{x} = (x_1, x_2, x_3, \ldots, x_l)$ and $\mathbf{y} = (y_1, y_2, y_3, \ldots, y_l)$, be two strings of l letters and define their interaction $U(\mathbf{x}, \mathbf{y})$ as

$$U(\mathbf{x}, \mathbf{y}) = \sum_{t=1}^{l} u(x_t, y_t). \tag{9}$$

The function $U(\mathbf{x}, \mathbf{y})$, as defined above, counts the number of unmatched letters between strings $\mathbf{x}$ and $\mathbf{y}$.

Introduce an 'inverse temperature' β and consider the Boltzmann factor $\mathrm{e}^{-\beta U}$. In the 'zero-temperature' limit we have

$$\lim_{\beta\to\infty} \mathrm{e}^{-\beta U(\mathbf{x},\mathbf{y})} = \begin{cases} 1, & \text{if } \mathbf{x} = \mathbf{y} \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

We see that the limit $\beta \to \infty$ is a 'no tolerance' limit [16], enforcing perfect matching of $\mathbf{x}$ and $\mathbf{y}$, i.e. $x_t = y_t, t = 1, 2, \ldots, l$. Let $\mathbf{y} = (y_1, y_2, \ldots, y_k)$ be a string of length $k \geqslant l$ and denoted by $\mathbf{y}_{a,l} = (y_{a+1}, y_{a+2}, \ldots, y_{a+l})$ the substring of length l starting at position $a, a = 0, 1, \ldots, k - l$. Furthermore let

$$f_a(\mathbf{x}, \mathbf{y}; \beta) = \mathrm{e}^{-\beta U(\mathbf{x},\mathbf{y}_{a,l})} \tag{11}$$

so that we have

$$f_a(\mathbf{x}, \mathbf{y}) \equiv \lim_{\beta\to\infty} f_a(\mathbf{x}, \mathbf{y}; \beta) = \begin{cases} 1, & \mathbf{x} = \mathbf{y}_{a,l} \\ 0, & \text{otherwise.} \end{cases} \tag{12}$$

Thus, $f_a(\mathbf{x}, \mathbf{y}) = 1$ if and only if $\mathbf{x}$ matches $\mathbf{y}$ at position a, and zero otherwise.

Likewise, let $f(\mathbf{x}, \mathbf{y})$ be a function that takes on the value one if the k-string $\mathbf{y}$ contains the given l-string $\mathbf{x}$ and zero otherwise. Note that the complement of the event that $\mathbf{x}$ matches $\mathbf{y}$ is the event that $\mathbf{x}$ does not match $\mathbf{y}$ anywhere. Thus, using equation (12), we can write

$$f(\mathbf{x}, \mathbf{y}) = 1 - \prod_{a=0}^{k-l} [1 - f_a(\mathbf{x}, \mathbf{y})]. \tag{13}$$

Letting $p(l, k; \mathbf{x})$ denote the probability that a randomly drawn k-string $\mathbf{y}$ contains a given l-string $\mathbf{x}$, we therefore find

$$p(l, k; \mathbf{x}) = 1 - \frac{1}{r^k} \sum_{\mathbf{y}} \prod_{a=0}^{k-l} [1 - f_a(\mathbf{x}, \mathbf{y})], \tag{14}$$

where r^k is the number of distinct k-strings of r-letters, and $\sum_{\mathbf{y}}$ denotes the sum over all such strings $\mathbf{y}$.

Generalizing the above equation to incorporate partial matches we obtain

$$p(l, k; \mathbf{x}) = \lim_{\beta \to \infty} p(l, k; \mathbf{x}, \beta), \tag{15}$$

where

$$p(l, k; \mathbf{x}, \beta) = 1 - \frac{1}{r^k} \sum_{\mathbf{y}} \prod_{a=0}^{k-l} [1 - f_a(\mathbf{x}, \mathbf{y}; \beta)]. \tag{16}$$

The products in equation (16) can be expanded and we obtain a Mayer-like sum

$$p(l, k; \mathbf{x}, \beta) = \frac{1}{r^k} \sum_{\mathbf{y}} \sum_{a} f_a - \frac{1}{r^k} \sum_{\mathbf{y}} \sum_{a<b} f_a f_b + \frac{1}{r^k} \sum_{\mathbf{y}} \sum_{a<b<c} f_a f_b f_c - \cdots, \tag{17}$$

which we can write as

$$p(l, k; \mathbf{x}, \beta) = \sum_{a} W^{(1)}(a; \mathbf{x}) - \sum_{a<b} W^{(2)}(a, b; \mathbf{x}) + \sum_{a<b<c} W^{(3)}(a, b, c; \mathbf{x}) - \cdots, \tag{18}$$

where

$$\begin{aligned} W^{(1)}(a; \mathbf{x}) &= \frac{1}{r^k} \sum_{\mathbf{y}} f_a(\mathbf{x}, \mathbf{y}; \beta) \\ W^{(2)}(a, b; \mathbf{x}) &= \frac{1}{r^k} \sum_{\mathbf{y}} f_a(\mathbf{x}, \mathbf{y}; \beta) f_b(\mathbf{x}, \mathbf{y}; \beta) \\ W^{(3)}(a, b, c; \mathbf{x}) &= \frac{1}{r^k} \sum_{\mathbf{y}} f_a(\mathbf{x}, \mathbf{y}; \beta) f_b(\mathbf{x}, \mathbf{y}; \beta) f_c(\mathbf{x}, \mathbf{y}; \beta) \\ &\cdots \end{aligned} \tag{19}$$

Using equations (8) and (11), we obtain

$$W^{(1)}(a; \mathbf{x}) = \frac{1}{r^l} [1 + (r-1)\,\mathrm{e}^{-\beta}]^l \equiv W^{(1)}. \tag{20}$$

Note that $W^{(1)}(a; \mathbf{x})$ is independent of a and $\mathbf{x}$.

Let us now turn to the second order term, $W^{(2)}(a, b; \mathbf{x})$ in equations (18) and (19). Here, we need to distinguish two cases: (i) $b - a \geqslant l$ and (ii) $b - a < l$.

In case (i), the set of indices of $\mathbf{y}_{a,l}$ and $\mathbf{y}_{b,l}$ are distinct and the evaluation of the partition sum proceeds analogously to equation (20) yielding

$$W^{(2)}(a, b; \mathbf{x}) = \left(\frac{1}{r^l}\right)^2 [1 + (r-1)\,\mathrm{e}^{-\beta}]^{2l}, \qquad |b - a| \geqslant l. \tag{21}$$

In case (ii), $|b-a|<l$, there is an overlap between the indices of $\mathbf{y}_{a,l}$ and $\mathbf{y}_{b,l}$. Letting $|b-a|=m$, we find

$$W^{(2)}(a,b;\mathbf{x}) = \frac{1}{r^{l+m}}[1+(r-1)\,\mathrm{e}^{-\beta}]^{2m}\prod_{t=1}^{l-m}[1+(r-1)\,\mathrm{e}^{-2\beta} \\ - u(x_t,x_{m+t})(1-\mathrm{e}^{-\beta})^2], \qquad |b-a|<l. \tag{22}$$

Note that $W^{(2)}(a,b;\mathbf{x})$, as defined in equations (21) and (22), depends on $\mathbf{x}$ only when $|b-a|<l$. Next, we perform the $\mathbf{x}$ average of $W^{(2)}(l,k;\mathbf{x})$,

$$\frac{1}{r^l}\sum_{\mathbf{x}} W^{(2)}(a,b;\mathbf{x}) = \frac{1}{r^{2l}}[1+(r-1)\,\mathrm{e}^{-\beta}]^{2l}. \tag{23}$$

The calculations leading to equations (22) and (23) are a little involved and can be found in the appendix.

Comparing equations (20) and (23), we see that once averaged over $\mathbf{x}$, $W^{(2)}$ factorizes as

$$W^{(2)} = \langle W^{(2)}(a,b;\mathbf{x})\rangle_x = (W^{(1)})^2, \tag{24}$$

or equivalently,

$$\langle f_a f_b\rangle_{y,x} = \langle f_a\rangle_{y,x}\langle f_b\rangle_{y,x}, \qquad a\neq b, \tag{25}$$

where, for simplicity, we have introduced the short-hand notation $\langle\cdots\rangle_{y,x}$ to denote averaging first over $\mathbf{y}$ and then $\mathbf{x}$.

Let us therefore make the approximation that all higher moments factorize similarly,

$$\left\langle f_{a_1} f_{a_2}\ldots f_{a_s}\right\rangle_{y,x} \simeq \left\langle f_{a_1}\right\rangle_{y,x}\left\langle f_{a_2}\right\rangle_{y,x}\cdots\left\langle f_{a_s}\right\rangle_{y,x}, \tag{26}$$

with $\{a_s\}$ being distinct. It can be readily shown that equation (26) is exact when $a_{i+1}-a_i>l$, i.e, there are no overlaps between the segments at position a_i. Upon substituting equation (26) into equation (17) and performing the $\mathbf{x}$ average we obtain the matching probability

$$p(l,k;\beta) = \langle p(l,k;\mathbf{x},\beta)\rangle_x, \tag{27}$$

with

$$p(l,k;\beta) = 1-\left(1-\frac{1}{r^l}[1+(r-1)\,\mathrm{e}^{-\beta}]^l\right)^{k-l+1}. \tag{28}$$

In the 'zero-temperature' limit ($\beta\to\infty$), this becomes

$$p(l,k) = 1-\left(1-\frac{1}{r^l}\right)^{k-l+1}. \tag{29}$$

For $r^l \gg k$, $p(l,k;\beta)$ has the asymptotic form

$$p(l,k;\beta) = 1-\exp\left(-\frac{k-l+1}{r^l}[1+(r-1)\,\mathrm{e}^{-\beta}]^l\right), \tag{30}$$

which for $\beta\to\infty$ becomes

$$p(l,k) = 1-\exp\left(-\frac{k-l+1}{r^l}\right). \tag{31}$$

For very large l this further reduces to

$$p(l,k) = \frac{k-l+1}{r^l}. \tag{32}$$

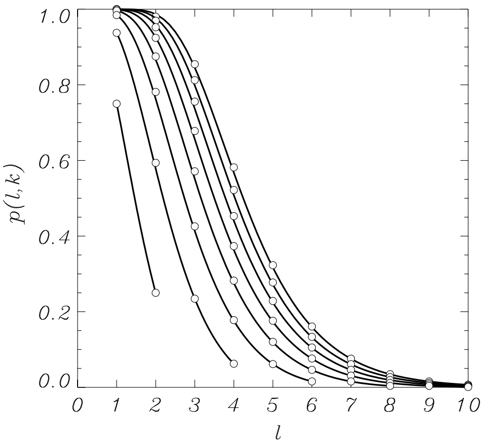


Figure 2. Comparison of the exact matching probability $p(k, l)$ (circles) with the approximate expression (29) (lines) for $r = 2$ and perfect matches. The curves are (from top to bottom) for values of $k = 16, 14, 12, 10, 8, 6, 4, 2$.

Note that a finite β acts like an enhanced matching probability, i.e., a false positive match. In the limit $\beta \to 0$, the matching probability becomes

$$\lim_{\beta\to 0} p(l, k; \beta) = 1. \tag{33}$$

Hence the 'high-temperature' limit of our model corresponds to indiscriminate matches.

Of course, the crucial approximation, equation (26), is not correct in general and one expects corrections coming from higher order correlations contained in equation (18). These correlations are due to the fact that if a given string $\mathbf{x}$ is matched at a position a, this affects the likelihood of matching the same string at any nearby location b with $|b - a| \lesssim l$. Nevertheless, the approximate result for $p(l, k)$, equation (29), is surprisingly good. Figure 2 shows a comparison of the matching probability obtained from exact enumeration carried out computationally, with the analytical expression (29) for $r = 2$ and perfect matches. As can be seen from the figure, there are only very small discrepancies for small l when $k > 2^l$, e.g. data points around $k = 16, 14, 12$ with $l = 4, 3, 2$. Since our expression for $p(l, k)$, equation (29), is exact for $l = 1$, there are no discrepancies at $l = 1$.

Note that equation (32) is the matching probability that can alternatively be obtained by assuming the probabilities of matching a string of length l at any position in a string of length k are independent and equal, $1/r^l$. Equation (29), on the other hand, is the matching probability that can also be found assuming the probabilities of *not* matching a string of length l at any position in a string of length k are independent and equal, $1 - 1/r^l$. Thus the factorization approximation, equation (26), leading to equation (28) implies that the probabilities of *not* matching at a given position are independent.

For the regime of interest, $k \lesssim 2^l$, this approximation leading to equation (29) is extremely good. We think that this is due to the fact that the factorization property underlying our approximation, equation (26), is exact for the two-point correlation function ($s = 2$), equation (24). This means that any corrections to this result must be coming from higher order correlations with strongly overlapping segments, since non-overlapping segments will factorize and thus reduce to lower order correlators. This is very similar to the connected cluster expansion in statistical mechanics [17]. Indeed, such an expansion can be set up; however, the calculations are rather tedious due to the discreteness of the problem and beyond the scope of this paper. Yet it is clear that the weight of an s-point correlation function with s overlapping (connected) segments must be very small for large s, since the overlap imposes very strong conditions on the structure of the string $\mathbf{x}$ to be matched.

For the remainder of the paper it is convenient to define the quantities t and z as

$$t = 1 - \frac{1}{r^l}[1 + (r-1)\,\mathrm{e}^{-\beta}]^l = 1 - z^l \tag{34}$$

$$z = \frac{1}{r}[1 + (r-1)\,\mathrm{e}^{-\beta}], \tag{35}$$

where we have suppressed the β, r and l dependence for clarity. Note that the effect of the number of letters in the alphabet r and the extent of mismatch as parameterized by the 'inverse temperature' β enter into the expression for $p(l, k; \beta)$ as a single parameter, z, as defined above. With the above definitions, equation (28) becomes

$$p(l, k; z) = 1 - t^{k-l+1} = 1 - (1 - z^l)^{k-l+1}. \tag{36}$$

The 'zero-temperature' limit is given by $z = r^{-1}$, while the 'high-temperature' limit is $z = 1$. The range of z is therefore, $z \in (r^{-1}, 1)$, which for $r \gg 1$, approaches $\in (0, 1)$.

We note in passing that the matching probability computed in this section is in the sense complementary to the problem of sequence alignment [18, 19], which has important applications in the study of proteins and DNA. The problem there is to identify subsequences of arbitrary length, showing strong similarity beyond pure statistical chance, *within* two long sequences sampling the same alphabet, possibly with different native probabilities. The pioneering work of Altschul, Karlin *et al* [18, 19] yields a probability distribution for the similarity *score* of such likely regions, under the assumption that the region with the highest score is unique (i.e., non-degenerate), that the two sequences searched are of comparable length, and sufficiently long. The scoring scheme is to a large extent arbitrary as long as the scores corresponding to some degree of matching are rare (and positive) while those corresponding to mismatches are much more probable (and negative). This arbitrariness may be removed by proper normalization and scores obtained via different schemes can be compared in a meaningful way. The matching probability computed in the present paper could be related to the probability for the highest score (corresponding to an exact match without gaps), holding for the entire length of the shorter sequence. However, our calculation makes no assumptions regarding the relative lengths of the two sequences, apart from the obvious requirement that $l \leqslant k$. The approximation to which we have to resort in the final solution works best when either the two sequences are almost of the same length, or if $k \lesssim r^l$. Moreover, there is no assumption regarding the number of times the highest score is achieved. More interestingly, the statistics of multiple high-scoring segments [20] could have been related to the out-degree statistics of a given node *had we taken each high-scoring match in the complete random sequence to correspond to a different edge*. As it is, a single edge corresponds to the presence of one or more occurrences of a shorter string, say G_i, inside a longer string G_j. That is, multiple occurrences of the shorter string within a subsequence of the complete random sequence are bunched together to result in a single edge between the nodes i and j.

We now turn to the calculation of the in- and out-degree distributions.

4. The degree distributions

In section 2 we showed that the subsequences $\{G_i\}$ of a random sequence C generate a network whose nodes are associated with these strings, and whose edges are defined by the matching relation equation (5). In this section we will derive the in- and out-degree distributions associated with this network.

Consider a randomly selected string G_i. The in- and out-degree of the corresponding nodes, $d_{\text{in}}(i)$ and $d_{\text{out}}(i)$, are defined by the total number of edges terminating in and originating from that node, respectively,

$$d_{\text{in}}(i) = \sum_j w_{ji} \qquad d_{\text{out}}(i) = \sum_j w_{ij}. \tag{37}$$

The corresponding in- and out-degree distributions are given by

$$n_{\text{in}}(d) = \sum_i \delta(d - d_{\text{in}}(i)) \qquad n_{\text{out}}(d) = \sum_i \delta(d - d_{\text{out}}(i)). \tag{38}$$

4.1. The out-degree distribution

Letting $\mathcal{G}_l$ denote the set of strings of length l, we can rewrite the out-degree distribution equation (38) as

$$n_{\text{out}}(d) = \sum_{l=1}^{L} n_l \left[\frac{1}{n_l} \sum_{j \in \mathcal{G}_l} \delta(d - d_{\text{out}}(j)) \right]. \tag{39}$$

For large n_l, the quantity in parentheses will approach the (conditional) probability $P_{\text{out}}(X_l = d|l)$ that a randomly selected string whose length is given to be l has an out-degree d.

In the limit $L, N \to \infty$, such that $N/L = p$, the ratio of the number of strings, N, to the length of the whole random sequence, L, remains constant, all the possible r^l realizations of random words of a given length l will be present with equal respective weights and we have

$$\lim_{N\to\infty} \frac{1}{n_l n_k} \sum_{i \in \mathcal{G}_l} \sum_{j \in \mathcal{G}_k} w_{ij} = p(l, k). \tag{40}$$

We will refer to this limit as the large-L limit.

The quantity $p(l, k)$, as defined in the above equation, is the probability that a randomly selected string of given length l matches another independently and randomly selected string of length k. This probability has been calculated in section 3 for the general case of imperfect matches, equation (28), as well as perfect matches, equation (29). Equations (39) and (40) show the self-averaging property of the degree distribution in the large-L limit.

Define the random variable X_{lk} as the number of edges originating from a randomly selected string of length l that terminate in strings of length k. Then X_l can be written as a sum of the random variables X_{lk},

$$X_l = \sum_{k \geqslant l} X_{lk}. \tag{41}$$

We can therefore write $\langle X_l \rangle$ as

$$\langle X_l \rangle = \frac{1}{n_l} \sum_{i \in \mathcal{G}_l} \sum_{k=l}^{L} \sum_{j \in \mathcal{G}_k} w_{ij}, \tag{42}$$

or,

$$\langle X_l \rangle = \sum_{k=l}^{L} n_k \left[\frac{1}{n_l n_k} \sum_{i \in \mathcal{G}_l} \sum_{j \in \mathcal{G}_k} w_{ij} \right], \tag{43}$$

where $\langle\cdots\rangle$ denotes an average over all the strings of length l in the complete random sequence, and we have in going from equation (41) to equation (43) used the weak law of large numbers under the assumption that L is large.

We see from equations (43) and (40) that in the large-L limit

$$\langle X_{lk}\rangle = n_k p(l,k), \tag{44}$$

and

$$\langle X_l\rangle = \sum_{k=l}^{L} n_k p(l,k). \tag{45}$$

Note that in the large-L limit X_{lk} is binomially distributed,

$$P(X_{lk} = d|l) = \binom{n_k}{d} p(l,k)^d (1-p(l,k))^{n_k-d}. \tag{46}$$

As can be seen from equation (41), X_l is a sum of the random variables X_{lk}, and thus in the large-L limit the central limit theorem assures that the distribution for X_l will approach a Gaussian distribution,

$$P_{\text{out}}(X_l = d|l) = \frac{1}{\sqrt{2\pi}\sigma_l}\exp\left[-\frac{(d-d_l)^2}{2\sigma_l^2}\right], \tag{47}$$

whose mean d_l and standard deviation σ_l are given by those of X_{lk}, equation (41), according to

$$d_l = \langle X_l\rangle = \sum_{k\geqslant l}\langle X_{lk}\rangle \tag{48}$$

$$\sigma_l^2 = \langle X_l^2\rangle - \langle X_l\rangle^2 = \sum_{k\geqslant l}\langle\sigma_{lk}^2\rangle, \tag{49}$$

where

$$\sigma_{lk}^2 = \langle X_{lk}^2\rangle - \langle X_{lk}\rangle^2. \tag{50}$$

For binomially distributed X_{lk} we have

$$\langle X_{lk}\rangle = n_k p(l,k) \tag{51}$$

$$\sigma_{lk}^2 = n_k p(l,k)(1-p(l,k)). \tag{52}$$

Using equation (36), one can readily carry out the sums in equations (48) and (49) to find

$$d_l = \frac{N}{p+qz^l}(qz)^l \tag{53}$$

$$\sigma_l^2 = d_l\frac{pt}{1-qt^2}. \tag{54}$$

Noting also that the probability of selecting a string of length l is pq^l, the total out-degree distribution is given by

$$P_{\text{out}}(d) = \sum_{l=1}^{L} pq^l P_{\text{out}}(X_l = d|l), \tag{55}$$

and thus in the large-L limit we obtain

$$P_{\text{out}}(d) = \sum_{l=1}^{L} pq^l \frac{1}{\sqrt{2\pi}\sigma_l} \exp\left[-\frac{(d-d_l)^2}{2\sigma_l^2}\right] \tag{56}$$

with d_l and σ_l given by equations (53) and (54), respectively.

As l becomes large, $p(l,k)$ decreases towards zero. Thus with increasing l the binomial distribution of X_{lk}, equation (46), will approach a Poisson distribution of the same mean. Note that the sum of independent and Poisson distributed random variables is also Poisson distributed with mean equal to the sum of the individual means. Thus for large l, X_l as defined in equation (41) is Poisson. For a Poisson distributed random variable the variance equals to its mean so that for large l we expect

$$\sigma_l^2 = d_l, \tag{57}$$

as can also be directly verified by taking the appropriate limit in equation (54).

Before proceeding we would like to briefly mention the size dependence of the node degrees and the number of edges of the resulting network. The out-degree d of a node is an extensive quantity with respect to the number of strings N, or equivalently the length of the random string L, so that we have $d \sim N \sim L$, as is readily seen from equation (53). Using equation (53), the average number of edges N_e in the resulting network can be readily calculated as

$$N_e = \sum_{l=1}^{L} n_l d_l = N^2 \frac{q^2 z}{1-q^2 z}, \tag{58}$$

where in the last step we have taken the upper limit of the sum to be infinity, since L is assumed to be very large.

4.1.1. Ensemble averages and finite size effects. The numerical data of [1] have been obtained from averaging over 500 realizations of a random sequence of length $L = 15\,000$ with $N = 750$. A finite sample size will cause sample to sample fluctuations in the number of strings, or 'random words.' An average over a large ensemble of different realizations will yield the same average values for the out-degrees as those obtained from a single random sequence of infinite length. However, averaging over many realizations will increase the fluctuations around the mean. It is not hard to see that this will affect predominantly nodes with large out-degrees (short strings) where there is already self-averaging within the random sequence, but with a distribution which varies from sample to sample.

Nodes with small out-degrees (long strings) correspond to rare matches and thus for these nodes there is no self-averaging within the sample. To see this, consider the extreme case, where a sample contains on average one or less matches for such a node. When an ensemble average is taken, the dominant contribution to the variance of the out-degree will come from the sample to sample fluctuations.

Denoting the mean and variance of the out-degree of a node of length l, that has been corrected for the finite size, by $\tilde{d}_l$ and $\tilde{\sigma}_l^2$, respectively, we have

$$\tilde{d}_l = d_l, \qquad \tilde{\sigma}_l^2 \to \sigma_l^2 \quad \text{for large } l. \tag{59}$$

In what follows we will re-calculate previously introduced statistics, taking into account the fluctuations in n_k. In order to avoid confusion, these quantities will be denoted with a tilde.

We can estimate $\tilde{\sigma}_l^2$ as follows. The random variable $\tilde{X}_{lk}$ itself is a sum of random variables:

$$\tilde{X}_{lk} = \sum_{j \in \mathcal{G}_k} Y_{lj}, \tag{60}$$

where $Y_{lj} = 1$ if the string G_j of length k matches the (given) string of length l and zero otherwise. Such an event constitutes a Bernoulli trial and its probability is $p(l, k)$. The mean and variance of Y_{lj} are given by

$$\langle Y_{lj}\rangle = p(l, k) \tag{61}$$

$$\left\langle Y_{lj}^2\right\rangle - \langle Y_{lj}\rangle^2 = p(l, k)(1 - p(l, k)). \tag{62}$$

The number of such trials is $\tilde{n}_k$, the number of elements of $\mathcal{G}_k$, and hence $\tilde{n}_k$ itself is a random variable. For sufficiently large N and for values of $\tilde{n}_k$ near the mean, the constraints, equation (4), can be neglected and the probability of finding $\tilde{n}_k$ strings of length k is approximately binomially distributed

$$P(\tilde{n}_k = n) = \binom{N}{n} (pq^k)^n (1 - pq^k)^{N-n}. \tag{63}$$

We thus find

$$\langle \tilde{n}_k\rangle = Npq^k \tag{64}$$

$$\tilde{\sigma}_{n_k}^2 = Npq^k(1 - pq^k). \tag{65}$$

Finding the distribution of a sum over a *finite* random number n of independently distributed random variables Y can be readily worked out using moment generating functions (see for example Feller [21]). In the case when both $\tilde{n}_k$ and Y_{lj} are binomial it turns out that the resulting distribution is binomial again, and we find

$$P(\tilde{X}_{lk} = d|l) = \binom{N}{d} [pq^k p(l, k)]^d [1 - pq^k p(l, k)]^{N-d}, \tag{66}$$

with mean and variance

$$\langle \tilde{X}_{lk}\rangle = Npq^k p(l, k) \tag{67}$$

$$\tilde{\sigma}_{lk}^2 = Npq^k p(l, k)[1 - pq^k p(l, k)]. \tag{68}$$

Thus, equation (66) is the finite-size result replacing equation (46), which is valid in the large-L limit. As remarked before, the means of the two distributions in equations (48) and (67) are equal, i.e., $\langle \tilde{X}_{lk}\rangle = \langle X_{lk}\rangle$. However, the variances are different and $\sigma_{lk}^2 < \tilde{\sigma}_{lk}^2$. Note that the second term in equation (68) is of the order of $(1 - p) \approx 1$ for small p. Thus we find to order p

$$\tilde{\sigma}_{lk}^2 = \langle \tilde{X}_{lk}\rangle, \tag{69}$$

and consequently to this order the mean and variance of $\tilde{X}_l$ become

$$\tilde{\sigma}_l^2 = \langle \tilde{X}_l\rangle = \langle X_l\rangle = d_l, \tag{70}$$

where d_l is the same mean out-degree that was previously obtained in the large L-limit, equation (53). The out-degree distribution corrected for finite-size effects thus becomes (c.f., equation (56))

$$\tilde{P}_{\text{out}}(d) = \sum_{l=1}^{L} pq^l \frac{1}{\sqrt{2\pi d_l}} \exp\left[-\frac{(d - d_l)^2}{2d_l}\right]. \tag{71}$$

Comparing this expression with the distribution obtained in the large-L limit, equation (56), we find that finite-size corrections are only present for small l, since we have already shown that

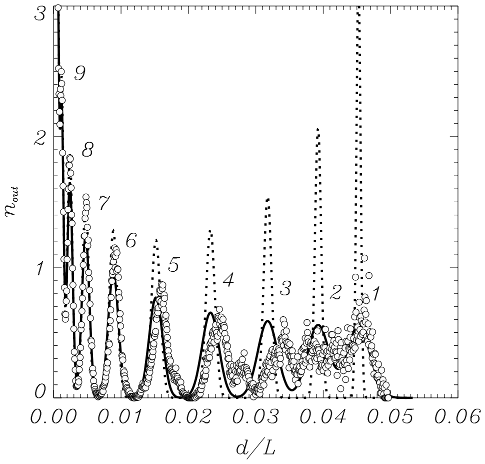


Figure 3. Comparison of the theoretical out-degree distributions with numerical data (circles) obtained for $L = 15\,000$. The dotted line shows the theoretical result for a network in the large L-limit, where the network is self-averaging and thus all possible realizations of a string of a given length l can be found. The number to the right of each peak refers to the node length l that contributes predominantly to that peak. The solid line is obtained after correcting for finite-size effects (see the text for details). In both cases, the locations of the peaks are accurately predicted. Note that the results for the large-L limit differ strongly in their predictions for the width and height of each peak for small l (large d). It is evident that the numerical data exhibit finite-size effects for short nodes, $l \lesssim 6$.

the relation $d_l = \sigma_l^2$ is also valid (viz. equation (57)) in the large l region for the large-L case. Figure 3 shows a comparison of the numerically obtained out-degree distribution (circles) with the theoretical expressions with and without finite-size corrections. The solid line is the analytical result for the out-degree distribution, equation (71), that takes into account finite-size corrections, while the dotted line corresponds to the case where the network is assumed to be self-averaging, i.e., equation (56) is satisfied, and thus sample to sample fluctuations can be neglected. Note the large difference from the observed behaviour for $l < 6$, $(d/L > 0.013)$ in the height and broadness of the distributions, when finite-size effects are not taken into account. The agreement of the finite-size corrected distribution with the numerical data, on the other hand, is rather good, and we conclude that finite-size effects present in the numerical data for short nodes are satisfactorily accounted for.

The locations of the peaks, d_l, coincide very well with the numerical data and we find indeed that each peak corresponds to the out-degree of nodes of a given length l. The locations of the peaks decrease exponentially with increasing l. The labels next to each peak show the string lengths l contributing predominantly to that peak.

Our reasoning above already shows that the oscillatory part of the out-degree distribution is highly susceptible to finite-size effects. It turns out that these oscillations are less pronounced or completely absent when *single* finite-size realizations of the network are considered. In other words, these oscillations become apparent only when averaging over many finite-size realizations, as we have done in our analysis.

We turn next to a discussion of the scaling behaviour.

4.1.2. Scaling behaviour.

Our analysis shows that the out-degree distribution is a superposition of Gaussian peaks with mean d_l and a variance that depends on the strength of finite-size effects, as discussed in the previous section. For large values of d, (small l) these peaks are well separated and one can readily obtain the envelope for the peaks. From equation (71) we see that the height E_l of a peak centred at d_l is

$$E_l = \frac{Npq^l}{\sqrt{2\pi d_l}}. \tag{72}$$

Using equation (53), we obtain the scaling behaviour

$$E(d) \approx d^{-\gamma_2} \tag{73}$$

with

$$\gamma_2 = \frac{1}{2}\frac{\ln z - \ln q}{\ln z + \ln q}. \tag{74}$$

For the bit string model with exact matches, i.e., for $r = 2$ and in the $\beta \to \infty$ limit, we find

$$\gamma_2 = \frac{1}{2}\frac{\ln 2 + \ln q}{\ln 2 - \ln q}. \tag{75}$$

For the numerical data shown, $q = 0.95$, yielding $\gamma_2 = 0.43$.

For smaller values of d (large l), the analysis presented above ceases to be valid, since the peaks start to overlap. In this regime, the contributions to the out-degree distributions come predominantly from matches between long strings which are rare. As was remarked previously, in this regime the distribution of $\tilde{X}_l$ will be Poisson, so that we have

$$p(d|l) = \frac{d_l^d}{d!}\,\mathrm{e}^{-d_l}, \tag{76}$$

with d_l as given before in (53). The out-degree distribution for small d is thus given by

$$p(d) = \sum_{l=l^*}^{\infty} pq^l \frac{d_l^d}{d!}\,\mathrm{e}^{-d_l}. \tag{77}$$

Since for small l the d_l values are quite large, the contributions from the small l terms will be suppressed heavily by the exponential factor, and therefore moving the cutoff l^* in the above sum down to 1 will not change the result of the summation significantly. Noting that for large l

$$d_l = \frac{N}{p}(qz)^l, \tag{78}$$

we see that d_l and $\Delta d_l = d_{l+1} - d_l$ approach zero in a geometric fashion. Thus the summation over l in equation (77) can be converted into an integration over $x = d_l$ with $\Delta x = d_l - d_{l+1}$ and we obtain

$$p(d) = \frac{c}{d!}\int_0^{x^*} x^{d-\gamma_2-\frac{1}{2}}\,\mathrm{e}^{-x}\,\mathrm{d}x, \tag{79}$$

where $x^* = d_{l^*}$ and c is an overall numerical constant,

$$c = \frac{p}{\ln qz}\left(\frac{N}{p}\right)^{-\frac{1}{2}-\gamma_2}. \tag{80}$$

The dominant contribution to the integrand comes from $x \approx d < x^*$ and we therefore extend the upper limit to infinity obtaining

$$p(d) = c\frac{\Gamma\left(d + \frac{1}{2} - \gamma_2\right)}{\Gamma(d+1)}, \tag{81}$$

where $\Gamma(x)$ is the gamma function. The leading order behaviour of $\ln\Gamma(x)$ is given asymptotically, for large$x$, by

$$\ln\Gamma(x) = \left(x - \frac{1}{2}\right)\ln x - x + \frac{1}{2}\ln 2\pi + O\left(\frac{1}{x}\right). \tag{82}$$

Using the above expansion, we obtain after a little algebra

$$\ln p(d) = \text{const.} - \left(\gamma_2 + \frac{1}{2}\right)\ln d + O\left(\frac{1}{d}\right). \tag{83}$$

It can be readily checked that this approximation for $\ln p(d)$ is good even for small values of d and thus $p(d)$ exhibits scaling behaviour, $p(d) \approx d^{-\gamma_1}$, with scaling exponent

$$\gamma_1 = \tfrac{1}{2} + \gamma_2. \tag{84}$$

For the numerical data with $z = 1/2$ and $q = 0.95$ we find $\gamma_1 = 0.93$.

As we have pointed out above, the crossover between the two scaling regimes occurs when the depression (minimum) between consecutive peaks disappears. This occurs roughly when

$$d_{l+1} + \frac{1}{\sqrt{2d_{l+1}}} > d_l - \frac{1}{\sqrt{2d_l}} \tag{85}$$

yielding, via equation (53),

$$d_l > \frac{1}{2}\frac{1}{1 - \sqrt{1 - qz}}. \tag{86}$$

For the values of the parameters employed in the numerical simulations, this gives $d_l > 6.59$, $\ln d_l > 1.9$, which is consistent with the data shown in figure 1.

We can also infer the large r behaviour of γ_1 and γ_2 for perfect matches. This corresponds to the case $z = 1/r$, equation (35). We find

$$\gamma_2 = \frac{1}{2}\frac{\ln r + \ln q}{\ln r - \ln q}, \tag{87}$$

and hence

$$\lim_{r\to\infty} \gamma_2 = \tfrac{1}{2} \tag{88}$$

and correspondingly $1/2 + \gamma_2 = \gamma_1 \to 1$ in this limit. Thus, as the number of letters in the alphabet is increased, the scaling exponents γ_1 and γ_2 approach the values 1 and $1/2$, respectively. Comparing with the values for $r = 2$, we see that the dependence of γ_1 and γ_2 on r, the number of letters in the alphabet, is rather weak.

4.2. The in-degree distribution

Consider a randomly selected string G_i of length l. Then the random variable X_{kl} that was introduced before counts the number of edges originating from a string of length $k \leqslant l$ and terminating in G_i. Thus, the in-degree of G_i is given by

$$X_{\text{in},l} = \sum_{k\leqslant l} X_{kl}. \tag{89}$$

The statistics of X_{kl} and hence of $X_{in,l}$ has been already obtained before and we find in the large-L limit,

$$d_{\text{in},l} = \sum_{k\leqslant l} n_k p(k,l) \tag{90}$$

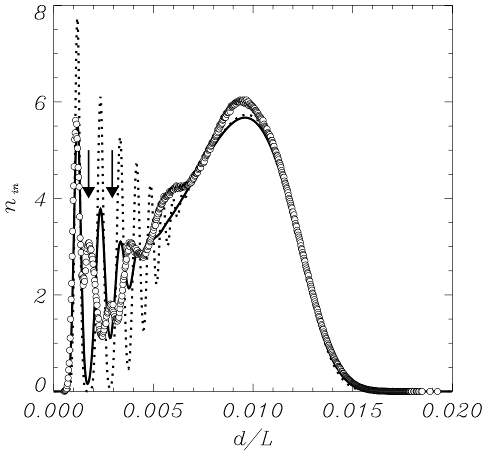


Figure 4. Comparison of the theoretical in-degree distributions with numerical data (circles) for $L = 40\,000$ and averaged over 20 000 realizations of the random string. The dotted line shows the theoretical result for a network in the large L-limit, where the network is self-averaging and thus almost all possible realizations of a string of a given length l can be found. The solid line is obtained after correcting for finite-size effects (see the text for further details).

$$\sigma^2_{\text{in},l} = \sum_{k\leqslant l} n_k p(k,l)(1-p(k,l)). \tag{91}$$

Noting also that the probability of selecting a string of length l is pq^l, the total in-degree distribution in the large-L limit is given by

$$P_{\text{in}}(d) = \sum_{l=1}^{L} pq^l \frac{1}{\sqrt{2\pi}\sigma_{\text{in},l}} \exp\left[-\frac{(d-d_{\text{in},l})^2}{2\sigma^2_{\text{in},l}}\right]. \tag{92}$$

When taking into account finite-size effects, the in-degree distribution becomes (cf section IV.A.1)

$$\tilde{P}_{\text{in}}(d) = \sum_{l=1}^{L} pq^l \frac{1}{\sqrt{2\pi}\tilde{\sigma}_{\text{in},l}} \exp\left[-\frac{(d-d_{\text{in},l})^2}{2\tilde{\sigma}^2_{\text{in},l}}\right], \tag{93}$$

where

$$\tilde{\sigma}^2_{\text{in},l} = \sum_{k\leqslant l} Npq^k p(k,l)(1-pq^k p(k,l)). \tag{94}$$

Note again that the average in-degree of a node is an extensive quantity with respect to the number of strings N, or equivalently the length of the random string L, so that we have $d \sim N \sim L$, as is readily seen from equations (89) and (90).

Unfortunately, we have not been able to obtain closed-form expressions for $d_{\text{in},l}$ and $\tilde{\sigma}_{\text{in},l}$, in a manner analogous to the expressions for the out-degree, equations (53) and (70). In the case of the in-degree distributions, equation (89) requires a sum over the first argument of the matching probability, $p(\cdot,\cdot;z)$, equation (36), rather than the second argument, as was the case for the out-degree distribution. Due to the complicated dependence of the matching probability on its first argument this sum is, as far as we can tell, intractable. The necessary summations were therefore carried out numerically.

Figure 4 shows a comparison of the two theoretical predictions, equations (92) and (93) with numerical data for $L = 40\,000$ averaged over 20 000 random realizations. Note the stark difference between the shape of the in- and out-degree distributions, figures 3 and 4. Apart from the distinct qualitative features, such as oscillatory behaviour for small d (rather than large d as in the out-degree distribution), the in-degree distribution is much narrower than the out-degree distribution.

The in-degree distribution equation (92), and its finite-size corrected form, equation (93), capture the qualitative features seen in the simulations. Inspecting figure 4, we see that the large degree behaviour agrees very well with the analytical form equation (92), the same is true for the very small degrees, but there are discrepancies for intermediate values. Note that the analytically predicted second peak in the low degree region of figure 4 is absent but instead we seem to have two neighboring peaks indicated by arrows in the figure.

Both of these peaks turn out to receive contributions from the in-degrees of strings of length 2. For a binary alphabet the possible values are 01, 10, 00 and 11. A node corresponding to strings of length 2 can have incoming edges only from identical strings or from strings of length 1. For each of the strings 01 and 01 the (average) in-degree is therefore equal to $d_{2,0} = n_1 + n_2/4$, where n_1 and n_2 are the average number of strings of length 1 and 2, respectively. For each of the strings 00 and 11 the average in-degree is $d_{2,1} = n_1/2 + n_2/4$, however. The total number of 2-strings falling into either of the two classes is $n_2/2$, so that one would expect two peaks in the in-distributions centred at $d_{2,0}$ and $d_{2,1}$ with respective masses $n_2/2$. The expected locations of these peaks have been indicated by the arrows in figure 4 and we see that they agree with the numerical findings. However, the actual situation is more involved than sketched above, because the locations of peaks due to other strings turn out to overlap. Thus for example, the two peaks indicated in figure 4 by arrows, receive also contributions from the in-degrees of nodes associated with other and longer strings. The same is also true for the other peaks in figure 4. Nonetheless, the above argument clearly shows that there is an additional fine structure in the in-degree distribution and that this structure is due to specific string contents.

5. Discussion

We have obtained analytical expressions for the in- and out-degree distributions of a contents-based network model which was introduced and studied numerically by Balcan and Erzan in [1]. We have shown that the behaviour of the out-degree distribution can be divided into two regimes: a short and putative scaling regime for small out-degrees that crosses over into an oscillatory regime for large out-degrees. An analytical expression for the crossover point has been obtained as well. We have found that the behaviour of the out-degree distribution for large out-degrees depends on the size of the network realizations from which the distribution was sampled. We have discussed these finite-size effects and have shown analytically how they effect the behaviour of the out-degree distribution.

Our results were obtained for a generalized class of contents-based network models in which a small number of imperfect matches (finite, but low, temperature) were allowed and strings were constructed from an alphabet of r letters. It turns out, however, that such generalizations do not alter the main numerical findings of the network model of Balcan and Erzan which involved a two-letter ($r = 2$) alphabet and perfect matches. The scaling behaviour which we have found, and even the numerical values of the leading scaling exponents γ_2 and $\gamma_1 = \gamma_2 + 0.5$ are robust under these generalizations. It should be noted that, in

$$\gamma_2 = \frac{1}{2}\frac{\ln z - \ln q}{\ln z + \ln q}, \tag{95}$$

we have $z \to 1/r$ for $\beta \to \infty$, while $z \to 1$ in the 'high-temperature' limit $\beta \to 0$, thus $r^{-1} \leqslant z \leqslant 1$. In the 'low-temperature,' or perfect matching, limit $\beta \to \infty$,

$$\gamma_2 \to (1/2)(1 - p/\ln r), \tag{96}$$

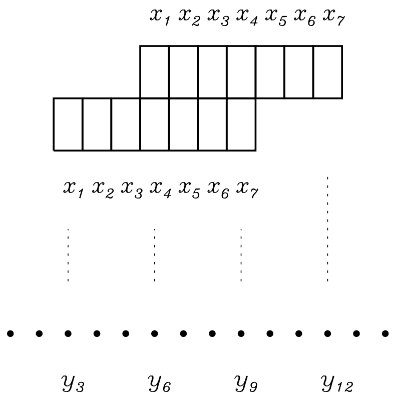
where p is a small number by assumption [1]. Even when allowing for a small number of mismatches, γ_2 depends very weakly on r and p. On the other hand, for either $r \to 1$, the trivial limit where no information is coded, or the high-temperature limit, where no matching conditions are satisfied, the scaling relation is altered qualitatively, with $\gamma_2 \to -1/2$.

The fine structure seen in the in-degree distribution is also present in the out-degree distribution [22], as is apparent from a closer inspection of the peaks in figure 3. It turns out that for the out-degree the additional structure arises from the fact that the probability of matching a given random string $\mathbf{x}$ of length l inside a random string of length k still depends, although only weakly, on the particular properties of the string $\mathbf{x}$ to be matched [23–25].

Since the in- as well as out-degrees are extensive quantities with respect to the number of strings N, or equivalently the length of the random string L, the oscillations in the degree distributions and the fine structure are resolvable only for sufficiently large string lengths L and number of realizations. Our treatment of the network topology explicitly ignores this dependency by considering the matching probability averaged over all possible strings $\mathbf{x}$ and can therefore be regarded as a coarse-grained description. The analysis presented can be readily extended to take this fine structure into account. However, this is beyond the scope of the current paper and will be left for future work.

It should be noted that in the present application, we have held the total length of the linear code, $L = \sum_l l\, n(l) + N$ constant, while N, the number of words (or equivalently, of delimiters) has been allowed to fluctuate. The single linear code with a special symbol (here the $(r+1)$th letter of the alphabet) indicating the beginning and ends of the 'words,' may easily be replaced by a collection of N strings coded by an alphabet of r letters, obeying an appropriately chosen length distribution, depending upon different applications. These different approaches are analogous to the use of the grand canonical as opposed to the canonical ensemble, with p corresponding to the fugacity. The use of a single linear random code emphasizes the spontaneous emergence of a nontrivial network of sequence matching interactions between substrings of the code of length L [18, 19]. It also lends itself to the elaboration of the evolutionary dynamics of this network, with the introduction of point mutations involving the random motion of the delimiters, besides random insertions, deletions and replacements of all elements of the $(r + 1)$-letter alphabet treated on an equal footing [26].

Finally let us remark that the present model can in fact be considered as a random network with different types of vertices indexed by l, having the probabilities $p(l, k)$ for the insertion of directed edges between nodes labelled by the ordered pair (l, k). The Poisson distribution of the out-degree for each node of type l is a hallmark of this underlying generalized Erdös-Renyi network. However, the oscillatory nature of the out-degree distribution for the total network, in the region of large degrees, depends on the precise geometric form of $p(l, k) \propto r^{-l}$, which, combined with the geometric form of $n(l)$, gives rise to the log-periodic form for the positions of these peaks, namely d_l, equation (53). In the regime where the peaks of the Poisson distributions are spaced more narrowly than their widths, namely, in the small d regime, the superposition of these Poisson peaks conspire to give rise to a power law on the average. The exponential l-dependence of $n(l)$ turns out to be algebraically tractable, but it may be conjectured that any distribution which has a tail that is decaying exponentially with l would give rise, all else remaining equal, to essentially the same scaling behaviour for the out-degree distribution in the large l (small d) regime, and therefore that $\gamma_1 \simeq 1$ has a high degree of universality.



x_1 x_2 x_3 x_4 x_5 x_6 x_7
x_1 x_2 x_3 x_4 x_5 x_6 x_7
y_3
y_6
y_9
y_{12}

Figure 5. A schematic representation of the y and x averages for the function $W^{(2)}(a, b; \beta)$ as defined in the text. The case shown in the figure corresponds to $l = 7$ with $a = 2$ and $b = 5$. The number of overlapping indices in the figure is $l - m(=4)$, with $m = b - a(=3)$. Because of the overlapping, when averaging over $\mathbf{x}$, these indices fall into $m(= 3)$ disjoint sets: $\{x_1, x_4, x_7\}$, $\{x_2, x_5\}$ and $\{x_3, x_6\}$.

Acknowledgments

One of us (AE) gratefully acknowledges partial support from the Turkish Academy of Sciences. MM gratefully acknowledges partial support from the Nahide and Mustafa Saydan Foundation. AK acknowledges support of FIRB01.

Appendix

Here we outline the calculations leading to equations (22) and (26). In section 3, we defined the function $W^{(2)}(a, b; \mathbf{x})$ as, equation (19),

$$W^{(2)}(a, b; \mathbf{x}) = \frac{1}{r^k} \sum_{\mathbf{y}} f_a(\mathbf{x}, \mathbf{y}; \beta) f_b(\mathbf{x}, \mathbf{y}; \beta). \tag{A.1}$$

As we pointed out in the text, when performing the sum over $\mathbf{y}$, two cases must be distinguished: (i) $|b - a| \geqslant l$ and (ii) $|b - a| < l$. In case (i), the set of indices of $\mathbf{y}_{a,l}$ and $\mathbf{y}_{b,l}$ are distinct and the evaluation of the partition sum proceeds in a manner analogous to equation (20) yielding

$$W^{(2)}(a, b; \mathbf{x}) = \left(\frac{1}{r^l}\right)^2 [1 + (r - 1)\,\mathrm{e}^{-\beta}]^{2l}, \qquad |b - a| \geqslant l. \tag{A.2}$$

In case (ii) there is an overlap between the indices of $\mathbf{y}_{a,l}$ and $\mathbf{y}_{b,l}$. Defining $|b - a| = m$, we find that there are $l - m$ overlapping indices, and thus there are $k - (l + m)$ distinct variables y_c that are neither in $\mathbf{y}_{a,l}$ nor in $\mathbf{y}_{b,l}$, so that a sum over the values of these indices will give $r^{k-(l+m)}$. Next, it is convenient to partition the remaining indices, $\{y_{a+1}, \ldots, y_{b+l}\}$, into the three disjoint sets, $S_1 = \{y_{a+1}, \ldots, y_{a+m}\}$, $S_2 = \{y_{a+m+1} = y_{b+1}, \ldots, y_{a+l} = y_{b+l-m+1}\}$ and $S_3 = \{y_{b+a-m+2}, \ldots, y_{b+l}\}$. Figure 5 shows an example for $l = 7$, with $a = 2$ and $b = 5$ along with the sets, $S_1 = \{y_3, y_4, y_5\}$, $S_2 = \{y_6, y_7, y_8, y_9\}$ and $S_3 = \{y_{10}, y_{11}, y_{12}\}$. With the definitions above, we find for $|b-a|<l$,

$$W^{(2)}(a,b;\mathbf{x}) = \frac{1}{r^{l+m}} \sum_{S_1} \exp\left(-\beta \sum_{t=1}^{k} u(x_t, y_{a+t})\right) \sum_{S_3} \exp\left(-\beta \sum_{t=1}^{m} u(x_{b+l-m+t}), y_{b+l-m+t}\right)$$
$$\times \sum_{S_2} \exp\left(-\beta \sum_{t=1}^{l-m} [u(x_t, y_{b+t}) + u(x_{m+t}, y_{b+t})]\right) \tag{A.3}$$

and carrying out the sums over the y variables, we obtain ($|b-a|<l$),

$$W^{(2)}(a,b;\mathbf{x}) = \frac{1}{r^{l+m}} [1+(r-1)\,\mathrm{e}^{-\beta}]^{2m} \prod_{t=1}^{l-m} [1+(r-1)\,\mathrm{e}^{-2\beta} - u(x_t, x_{t+m})(1-\mathrm{e}^{-\beta})^2]. \tag{A.4}$$

Next, it is useful to introduce the $r \times r$ matrix, $M(x,y)$, as

$$M(x,y) = 1+(r-1)\,\mathrm{e}^{-2\beta} - u(x,y)(1-\mathrm{e}^{-\beta})^2, \tag{A.5}$$

with $x, y \in \{0,1,2,\ldots,r-1\}$. From the properties of u, equation (8), we find that

$$M(x,y) = \begin{cases} 1+(r-1)\,\mathrm{e}^{-2\beta}, & x = y \\ (r-2)\,\mathrm{e}^{-2\beta} + 2\,\mathrm{e}^{-\beta}, & x \neq y \end{cases} \tag{A.6}$$

and equation (A.4) can therefore be written as

$$W^{(2)}(a,b;\mathbf{x}) = \frac{1}{r^{l+m}} [1+(r-1)\,\mathrm{e}^{-\beta}]^{2m} \prod_{t=1}^{l-m} M(x_t, x_{t+m}). \tag{A.7}$$

Proceeding to perform the average over $\mathbf{x}$,

$$W^{(2)}(a,b) = \frac{1}{r^l} \sum_{\mathbf{x}} W^{(2)}(a,b;\mathbf{x}), \tag{A.8}$$

observe that in equation (A.7) the variables $\mathbf{x}$ can be partitioned into k disjoint sets X with the additional property that if $x_t \in X$, by implication $x_{t+m} \in X$. The situation is shown schematically in figure (5) for $m=3$, where we have the 3 disjoint sets, $\{x_1, x_4, x_7\}$, $\{x_2, x_5\}$ and $\{x_3, x_6\}$. Denoting these sets as $X_1, X_2, \ldots X_m$, and their respective number of elements as $n_1, n_2, \ldots, n_m$ $(n_1 + n_2 + \cdots n_m = l)$, we see that the product in equation (A.7) can be factorized as

$$\prod_{t=1}^{l-m} M(x_t, x_{t+m}) = \prod_{x_t \in X_1} M(x_t, x_{t+m}) \cdots \prod_{x_t \in X_m} M(x_t, x_{t+m}). \tag{A.9}$$

Performing the summation over each of the factors we have for the first factor

$$\sum_{X_1} \prod_{x_t \in X_1} M(x_t, x_{t+m}). \tag{A.10}$$

It can be easily shown that the sum over the variables $x_t \in X_1$ reduces to an $n_1 - 1$ fold matrix product. Denoting the matrix elements of the matrix M^n by $(M^n)_{xy}$, we therefore find

$$\sum_{X_1} \prod_{x_t \in X_1} M(x_t, x_{t+m}) = \sum_{x,y} (M^{n_1-1})_{xy} \tag{A.11}$$

and hence

$$\frac{1}{r^l} \sum_{\mathbf{x}} \prod_{t=1}^{l-m} M(x_t, x_{t+m}) = \frac{1}{r^l} \prod_{s=1}^{m} \sum_{x,y} (M^{n_s-1})_{xy}. \tag{A.12}$$

Owing to the structure of the matrix M, equation (A.6), powers of M retain the same structure, as can be readily shown, and we therefore have

$$(M^n)_{(xy)} = \begin{cases} A_n, & x = y \\ B_n, & x \neq y. \end{cases} \tag{A.13}$$

The quantities A_n and B_n can be evaluated recursively, and one finds after a little algebra,

$$\begin{pmatrix} A_{n+1} \\ B_{n+1} \end{pmatrix} = Q_n \begin{pmatrix} A_1 \\ B_1 \end{pmatrix}, \tag{A.14}$$

where

$$Q_n = \frac{1}{r} \begin{pmatrix} (r-1)\lambda_+^n + \lambda_-^n & -(r-1)\lambda_+^n + (r-1)\lambda_-^n \\ -\lambda_+^n + \lambda_-^n & \lambda_+^n + (r-1)\lambda_-^n \end{pmatrix}, \tag{A.15}$$

and

$$\lambda_+ = [1 - \mathrm{e}^{-\beta}]^2 \tag{A.16}$$

$$\lambda_- = [1 + (r-1)\,\mathrm{e}^{-\beta}]^2. \tag{A.17}$$

We therefore find

$$\sum_{x,y} (M^n)_{xy} = rA_n + r(r-1)B_n, \tag{A.18}$$

and thus

$$\sum_{x,y} (M^n)_{xy} = r[1 + (r-1)\,\mathrm{e}^{-\beta}]^{2n}. \tag{A.19}$$

Substituting equation (A.19) back into equation (A.12) we have

$$\frac{1}{r^l} \sum_{\mathbf{x}} \prod_{t=1}^{l-m} M(x_t, x_{t+m}) = \frac{1}{r^l} \prod_{s=1}^{m} r[1 + (r-1)\,\mathrm{e}^{-\beta}]^{2(n_s-1)} \tag{A.20}$$

and noting that $n_1 + n_2 + \cdots n_m = l$, we finally obtain

$$\frac{1}{r^l} \sum_{\mathbf{x}} \prod_{t=1}^{l-m} M(x_t, x_{t+m}) = \frac{1}{r^l} r^m [1 + (r-1)\,\mathrm{e}^{-\beta}]^{2(l-m)}, \tag{A.21}$$

which when substituted into equations (A.8) and (A.7) yields the final result, equation (23),

$$\frac{1}{r^l} \sum_{\mathbf{x}} W^{(2)}(a, b; \mathbf{x}) = \frac{1}{r^{2l}} [1 + (r-1)\,\mathrm{e}^{-\beta}]^{2l}. \tag{A.22}$$

Note that we obtain the same result as for the case $|b - a| > l$, equation (A.2). In particular, we see that once averaged over $\mathbf{x}$, $W^{(2)}$ is independent of a and b.

References

[1] Balcan D and Erzan A 2004 *Eur. Phys. J.* B **38** 253

[2] Pastor-Satorras R and Vespignani A 2004 *Evolution and Structure of the Internet—A Statistical Physics Approach* (Cambridge: Cambridge University Press)

[3] Hannon G J 2002 *Nature* **418** 244

[4] He L and Hannon G J 2004 *Nature Rev. Genetics* **5** 522–31

[5] Sole R V and Pastor-Satorras R 2002 Complex networks in genomics and proteomics *Handbook of Graphs and Networks* ed S Bornholdt and H G Schuster (Berlin: Wiley-VCH)

[6] Alberts B *et al* 2002 *Molecular Biology of the Cell* (New York: Garland Science) ch 9

[7] Lee T I *et al* 2002 *Science* **298** 799

[8] Harbison C T *et al* 2004 *Nature* **431** 99

[9] Guelzim N, Bottani S, Bourgine P and Kepes F 2002 *Nature Genetics* **31** 60–3

[10] Dobrin R, Beg Q K, Barabasi A-L and Oltvai Z N 2004 *BMC Bioinformatics* **5** 1

[11] Tong A H Y *et al* 2004 *Science* **303** 808

[12] Perelson A S and Weisbuch G 1997 *Rev. Mod. Phys.* **69** 1219

[13] Erdös P and Renyi A 1959 *Publ. Mat. (Debrecen)* **6** 290
Erdös P and Renyi A 1960 *Publ. Mat. Inst. Hung. Acad. Sci.* **5** 17
Erdös P and Renyi A 1961 *Bull. Inst. Int. Stat.* **38** 343
Cited in Albert R and Barabasi A-L 2002 *Rev. Mod. Phys.* **74** 47

[14] Barabasi A-L and Albert R 1999 *Science* **286** 509

[15] Albert R and Barabasi A-L 2002 *Rev. Mod. Phys.* **74** 47

[16] Özçelik S and Erzan A 2003 *Int. J. Mod. Phys.* C **14** 169

[17] Uhlenbeck G E and Ford G W 1963 *Lectures in Statistical Mechanics* (Providence, RI: American Mathematical Society)
Huang K 1987 *Statistical Mechanics* (New York: Wiley)

[18] Altschul S, Gish W, Miller W, Meyers E W and Lipman D 1990 *J. Mol. Biol.* **225** 403

[19] Karlin S and Altschul S F 1990 *Proc. Natl Acad. Sci.* **87** 2264

[20] Karlin S and Altschul S F 1993 *Proc. Natl Acad. Sci.* **90** 5673

[21] Feller W 1971 *An Introduction to Probability Theory and its Applications* (New York: Wiley)

[22] Bilge A H, Balcan D and Erzan A 2004 The shift-match number and string matching probabilities for binary sequences *Preprint* q-bio.GN/0409023

[23] Guibas L J and Odlyzko A M 1978 *SIAM J. Appl. Math* **35** 401

[24] Guibas L J and Odlyzko A M 1981 *J. Comb. Theor.* **30A** 19

[25] Guibas L J and Odlyzko A M 1981 *J. Comb. Theor.* **30A** 183

[26] Sengun Y and Erzan A 2005 Content-based network model with duplication and divergence *Preprint* cond-mat/0510279