Largest Bipartite Sub-matchings of a Random Ordered Matching or a Problem with Socks

Andrzej Dudek[†], Jarosław Grytczuk[‡], and Andrzej Ruciński[§]

[†]*Department of Mathematics, Western Michigan University, Kalamazoo, MI, USA*
Email: andrzej.dudek@wmich.edu

[‡]*Faculty of Mathematics and Information Science, Warsaw University of Technology, Warsaw, Poland*
Email: jaroslaw.grytczuk@pw.edu.pl

[§]*Department of Discrete Mathematics, Adam Mickiewicz University, Poznań, Poland*
Email: rucinski@amu.edu.pl

ABSTRACT: Let M be an *ordered matching* of size n, that is, a partition of the set $[2n]$ into 2-element subsets. The *sock number* of M is the maximum size of a sub-matching of M in which all left-ends of the edges precede all the right-ends (such matchings are also called bipartite). The name of this parameter comes from an amusing "real-life" problem posed by Bosek, concerning an *on-line* pairing of randomly picked socks from a drying machine. Answering one of Bosek's questions we prove that the sock number of a random matching of size n is asymptotically equal to $n/2$. Moreover, we prove that the expected average number of socks waiting for their match during the whole process is equal to $\frac{2n+1}{6}$. Analogous results are obtained if socks come not in pairs, but in sets of size $r \geqslant 2$, which corresponds to a similar problem for random ordered r-matchings. We also attempt to enumerate matchings with a given sock number.

Keywords: Combinatorial enumeration; Ordered matchings; Random matchings
2020 Mathematics Subject Classification: 05C80; 05C30; 05C70

1. Introduction

Bartłomiej Bosek [3] has asked the following "practical" questions. There is a laundry sack with n different pairs of (clean) socks. You pull randomly socks from the sack, one by one, and place them on the floor, unless the sock in your hand matches one already on the floor in which case you bundle the pair up and put it into a drawer. What is the largest number of socks on the floor at any given time? What is the average number of socks on the floor throughout the whole process? Of course, these are questions about random variables in a suitable probability space.

The linear order in which the socks are pulled from the sack can be identified with one of $(2n)!/2^n$ permutations with n pairwise repetitions. However, as it really does not matter which socks are pulled and when, but only whether they form a pair with one already on the floor, we can truncate the space down to just all $(2n)!/(2^n n!)$ (ordered) n-element matchings of the set $[2n] = \{1, 2, \ldots, 2n\}$.

For instance, matching $\{1, 5\}, \{2, 3\}, \{4, 6\}$ corresponds to the order of socks in which the first and the fifth socks pulled are matched and so are the second and the third, as well as the fourth and the sixth. Here the situation on the floor goes through the following six stages: $1, 12, 1$ (because the second and the third sock are put together in the drawer), $14, 4, \varnothing$. So, in this case, the answers to the two above questions are 2 and $7/6$, respectively. If, instead, we are looking at matching $\{1, 5\}, \{2, 4\}, \{3, 6\}$, the answers are 3 and $3/2$.

There is a convenient way to represent ordered matchings in terms of *Gauss words*, i.e., words in which every letter of an n-element alphabet appears exactly twice and words obtained by permuting their letters are identified (e.g., $ABBA = BAAB$). Clearly, we get such a word from an ordered matching if the ends of each edge are replaced with a pair of identical letters, different from all other pairs of letters. For example, the two instances above can be encoded as $ABBCAC$ and $ABCBAC$, respectively.

Formally, let M be a matching on $[2n]$. For each $k \in [2n]$, let $x_k := x_k(M)$ be the number of edges of M with one endpoint in $[k]$ and the other in $[2n]\backslash[k]$. Note that x_k represents exactly the number of socks on the

floor after k socks have been pulled from the sack. For that reason we call the sequence $(x_k)_{k=1}^{2n}$ the *sockuence* of M (the term coined in by Martin Milanič).

Let $\mathbb{RM}(n)$ be a random matching on $[2n]$, that is, a matching picked uniformly at random out of the set of all $\frac{(2n)!}{(2)^n\,n!}$ matchings on the set $[2n]$. Thus, defining random variables $X_k = x_k(\mathbb{RM}(n))$, the two questions of Bosek ask about distributions of the random variables $Y = \max_k X_k$ – the largest value of the sockuence of $\mathbb{RM}(n)$ and $\bar{X} = \sum_k X_k/(2n)$ – the average value of the sockuence, respectively. In this paper we determine both random variables asymptotically, the latter being a simpler task.

Theorem 1.1. *We have* $\mathbb{E}\bar{X} = \frac{2n+1}{6}$ *and, for every* $\omega(n) \to \infty$, *a.a.s.*

$$|\bar{X} - n/3| \leqslant \omega(n)\sqrt{n}.$$

As for the former task, note that the random variable Y equals the size of the largest *bipartite* sub-matching of $\mathbb{RM}(n)$, that is, a sub-matching consisting of edges whose all left-ends precede all right-ends. Equivalently, a bipartite matching is one with the interval chromatic number equal to 2 (see, e.g., [8] for definition). In the word notation, a *bipartite* ordered matching is just a Gauss word whose second half is a permutation of the first half, as it happens, for instance, in the word $ABCDBDAC$. There is yet another alternative definition of bipartite matchings in terms of forbidden sub-structures. Indeed, there are three types, or patterns, in which two edges may intertwine: an *alignment* $AABB$, a *nesting* $ABBA$, or a *crossing* $ABAB$. Then a bipartite matching is precisely one with no alignments.

It has been known (see [10] for alignments, [2] for nestings and crossings, see also [6]) that for each of the three patterns, a.a.s. the largest size of a sub-matching of $\mathbb{RM}(n)$ with all pairs of edges forming that pattern has size $\Theta(\sqrt{n})$. It follows that a.a.s. $Y = \Omega(\sqrt{n})$. Here we show that, in fact, a.a.s. Y is a linear function of n.

Theorem 1.2. *There exists a constant* $C > 0$ *such that, a.a.s.,*

$$|Y - n/2| \leqslant C\sqrt{n \log n}.$$

(This has been proved in [12, Theorem 4.7] with an unspecified term $o(n)$ instead of our $\sqrt{n \log n}$.)

The problem can be generalized to r-matchings as follows. Imagine a distant planet inhabited by multi-leg creatures which for that reason wear not pairs but r-packs of socks, for some fixed $r \geqslant 2$. Then the process of placing them in a drawer differs only in that socks are taken up from the floor when the last one from a pack is pulled from the sack. This corresponds to a random r-matching $\mathbb{RM}^{(r)}(n)$ drawn uniformly at random from all $(rn)!/(r!^n n!)$ r-matchings (i.e., partitions into r-element subsets) of the set $[rn]$. Again, r-matchings can be represented as r-fold Gauss words, where each letter appears exactly r times and permuting the letters has no effect.

For instance, let $r = 3$ and $n = 12$ and look at the 4-edge 3-matching encoded by triples of distinct letters as $AABCDDDCBCBA$. Here the process goes as follows:

$$A, AA, AAB, AABC, AABCD, AABCDD, AABC, AABCC, AABCCB, AABB, AA, \varnothing,$$

so the maximum achieved (twice) is 6, while the average length is $42/12 = 7/2$.

Here again, we may express the problem in terms of bipartite sub-matchings with edges, however, bearing weights corresponding to the number of vertices "on the left". Formally, given an r-matching M on $[rn]$, for each $k \in [rn]$, let M_k be the set of all edges of M with nonempty intersections with both $[k]$ and $[rn]\backslash[k]$. Further, let $x_k(M) = \sum_{e \in M_k} |e \cap [k]|$. Note that $x_k(M)$ represents exactly the number of socks on the floor after k socks have been pulled from the sack. For instance, in the above example, M_8 consists of three edges, marked by letters A, B, and C, and $x_8(M) = 2 + 1 + 2 = 5$.

Let $\mathbb{RM}^{(r)}(n)$ be a random r-matching on $[rn]$ and define random variables

$$X_k^{(r)} = x_k(\mathbb{RM}^{(r)}(n)), \quad \overline{X^{(r)}} = \sum_k X_k^{(r)}/(rn), \quad \text{and} \quad Y^{(r)} = \max_k X_k^{(r)}.$$

We generalize Theorems 1.1 and 1.2 to r-matchings as follows.

Theorem 1.3. *We have* $\mathbb{E}\overline{X^{(r)}} = \frac{(r-1)(rn+1)}{2(r+1)}$ *and, for every* $\omega(n) \to \infty$, *a.a.s.*

$$\left|\overline{X^{(r)}} - \frac{(r-1)rn}{2(r+1)}\right| \leqslant \omega(n)\sqrt{n}.$$

Theorem 1.4. *For every* $r \geqslant 2$, *there exists a constant* $C > 0$ *such that, a.a.s.*

$$\left|Y^{(r)} - \frac{(r-1)n}{r^{\frac{1}{r-1}}}\right| \leqslant C\sqrt{n \log n}.$$

In particular, for $r = 3$, we infer that a.a.s. $\overline{X^{(3)}} \sim \frac{3}{4}n$ and $Y^{(3)} \sim \frac{2\sqrt{3}}{3}n$.

It is worth noting that Theorems 1.1 and 1.3 can be reformulated in terms of the distribution of the lengths of edges in a random matching. Indeed, given an edge $e = \{i, j\}$, where $1 \leqslant i < j \leqslant 2n$, let $\ell(e) := j - i$ be the *length* of e. Then, by a standard double-counting argument, for every matching M of $[2n]$,

$$\sum_{k=1}^{2n} x_k(M) = \sum_{e \in M} \ell(e).$$

Defining random variables $L := \sum_{e \in \mathbb{RM}(n)} \ell(e)$ and $\bar{L} := \frac{L}{n}$, we have $\bar{L} = 2\bar{X}$ and thus may restate Theorem 1.1 as follows.

Corollary 1.1. *We have* $\mathbb{E}\bar{L} = \frac{2n+1}{3}$ *and, for every* $\omega(n) \to \infty$, *a.a.s.*

$$|\bar{L} - 2n/3| \leqslant \omega(n)\sqrt{n}.$$

This connection can be generalized to arbitrary $r \geqslant 2$, provided one comes up with a suitable definition of the length of an r-element subset of $[rn]$ reflecting the nature of the socks problem, namely that the smaller vertices get higher weights (as they stay longer on the floor). To this end, given $e = \{i_1 < i_2 < \cdots < i_r\} \in \mathbb{RM}^{(r)}(n)$, let

$$\ell(e) := (i_2 - i_1) + 2(i_3 - i_2) + \cdots + (r-1)(i_r - i_{r-1}) = (i_r - i_1) + (i_r - i_2) + \cdots (i_r - i_{r-1}).$$

It is not difficult to see that, for every r-matching M of $[rn]$,

$$\sum_{k=1}^{rn} x_k(M) = \sum_{e \in M} \ell(e).$$

Indeed, vertex $i_j \in e$, $1 \leqslant j \leqslant r$, contributes 1 to the numbers $x_{i_j}, x_{i_j+1}, \ldots, x_{i_r-1}$ and 0 to all others, so its contribution equals $i_r - i_j$ and thus the total contribution of e toward $\sum_{k=1}^{rn} x_k(M)$ is precisely $\ell(e)$.

Defining random variables $L^{(r)} := \sum_{e \in \mathbb{RM}^{(r)}(n)} \ell(e)$ and $\bar{L}^{(r)} := \frac{L^{(r)}}{n}$, we have $\bar{L}^{(r)} = r\bar{X}^{(r)}$ and thus may restate Theorem 1.3 as follows.

Corollary 1.2. *We have* $\mathbb{E}\bar{L}^{(r)} = \frac{r(r-1)(rn+1)}{2(r+1)}$ *and, for every* $\omega(n) \to \infty$, *a.a.s.*

$$\left|\bar{L}^{(r)} - \frac{(r-1)r^2n}{2(r+1)}\right| \leqslant \omega(n)\sqrt{n}.$$

2. Proofs

In this section, we prove Theorems 1.3 and 1.4. Recall that $X_k^{(r)} = \sum_{e \in \mathbb{RM}_k^{(r)}(n)} |e \cap [k]|$, where $\mathbb{RM}_k^{(r)}(n)$ the set of all edges of $\mathbb{RM}^{(r)}(n)$ with nonempty intersections with both $[k]$ and $[rn]\backslash[k]$. So we may assume that $1 \leqslant k \leqslant rn - 1$.

Both proofs rely on high concentration of respective random variables around their means: $\overline{X^{(r)}}$ for Theorem 1.3, the $X_k^{(r)}$'s for Theorem 1.4. To this end, we will use the Azuma-Hoeffding inequality for random permutations (see, e.g., Lemma 11 in [7] or Section 3.2 in [11]). This is feasible because ordered matchings can be produced by generating random permutations. Indeed, let π be a permutation of $[rn]$. It can be chopped off into an r-matching $M_\pi := \{\pi(1) \ldots \pi(r), \pi(r+1) \ldots \pi(2r), \ldots, \pi(rn-r+1) \ldots \pi(rn)\}$ and, clearly, there are exactly $(r!)^n n!$ permutations π yielding the same matching. Thus, we can use the following lemma. By swapping two elements in a permutation π_1 we mean fixing two indices $i < j$ and creating a new permutation π_2 with $\pi_2(i) = \pi_1(j)$, $\pi_2(j) = \pi_1(i)$, and $\pi_2(\ell) = \pi_1(\ell)$ for all $\ell \neq i, j$. Let Π_N denote a permutation selected uniformly at random from all $N!$ permutations of $[N]$.

Lemma 2.1 ([7] or [11]). *Let* $h(\pi)$ *be a function defined on the set of all permutations of order* N *which satisfies the following Lipschitz-type condition: there exists a constant* $c > 0$ *such that whenever a permutation* π_2 *is obtained from a permutation* π_1 *by swapping two elements, we have* $|h(\pi_1) - h(\pi_2)| \leqslant c$. *Then, for every* $\eta > 0$,

$$\mathbb{P}(|h(\Pi_N) - \mathbb{E}[h(\Pi_N)]| \geqslant \eta) \leqslant 2\exp\{-2\eta^2/(c^2N)\}.$$

With respect to the Lipschitz condition, the following observation is crucial for us. Let π_1 be a permutation of $[rn]$ and permutation π_2 be obtained from π_1 by swapping two elements. This way we can destroy (or create) at most two edges which may jointly contribute at most $2(r-1)$ to $X_k^{(r)}$. Hence, for each k the random variable $X_k^{(r)}$ satisfies the Lipschitz condition with $c = 2(r-1)$ and so does $\overline{X^{(r)}}$.

Equipped with the above concentration tool, the next thing we need is a formula for $\mathbb{E}(X_k^{(r)})$.

Claim 2.1. *For all* $1 \leqslant k \leqslant rn - 1$,

$$\mathbb{E}(X_k^{(r)}) = k - \frac{r\binom{k}{r}}{\binom{rn-1}{r-1}}.$$

Proof of Claim 2.1. In order to find the expected value of $X_k^{(r)}$, fix an integer $1 \leqslant j \leqslant r - 1$ and observe that there are $\binom{k}{j}\binom{rn-k}{r-j}$ possible edges having exactly j vertices in $[k]$. Since a fixed edge appears in a randomly chosen matching with probability $1/\binom{rn-1}{r-1}$, we get

$$\mathbb{E}(X_k^{(r)}) = \sum_{j=1}^{r-1} \frac{j\binom{k}{j}\binom{rn-k}{r-j}}{\binom{rn-1}{r-1}} = \frac{k\sum_{j=0}^{r-1}\binom{k-1}{j}\binom{rn-k}{r-1-j} - k\binom{k-1}{r-1}}{\binom{rn-1}{r-1}} = k - \frac{r\binom{k}{r}}{\binom{rn-1}{r-1}}.$$

$\square$

Proof of Theorem 1.3. By Claim 2.1 and the linearity of expectation

$$\mathbb{E}(\overline{X^{(r)}}) = \frac{1}{rn}\sum_{k=1}^{rn-1}\mathbb{E}(X_k^{(r)}) = \frac{1}{rn}\sum_{k=1}^{rn-1}\left(k - \frac{r\binom{k}{r}}{\binom{rn-1}{r-1}}\right) = \frac{rn-1}{2} - \frac{1}{n\binom{rn-1}{r-1}}\sum_{k=1}^{rn-1}\binom{k}{r}$$

$$= \frac{rn-1}{2} - \frac{1}{n\binom{rn-1}{r-1}}\binom{rn}{r+1} = \frac{rn-1}{2} - \frac{rn-r}{r+1} = \frac{(r-1)(rn+1)}{2(r+1)}.$$

Recall that $\overline{X^{(r)}}$ satisfies the Lipschitz condition with $c = 2(r-1)$. Hence, given $\omega(n) \to \infty$, Lemma 2.1 applied with $N = rn$, $h(\pi) = \sum_{k=1}^{rn} x_k(M_\pi)/(rn)$, $c = 2(r-1)$, and $\eta = \frac{1}{2}\omega(n)\sqrt{n}$ yields (for large n)

$$\mathbb{P}\left(\left|\overline{X^{(r)}} - \frac{(r-1)rn}{2(r+1)}\right| \geqslant \omega(n)\sqrt{n}\right) \leqslant \mathbb{P}\left(\left|\overline{X^{(r)}} - \mathbb{E}\overline{X^{(r)}}\right| \geqslant \frac{1}{2}\omega(n)\sqrt{n}\right)$$

$$\leqslant 2\exp\left\{-\frac{\omega^2(n)}{2c^2r}\right\} = o(1).$$

$\square$

Proof of Theorem 1.4. This proof is more tricky, as we do not have an exact formula for $\mathbb{E}Y$, where recall $Y = \max_k X_k^{(r)}$. Instead, we apply Lemma 2.1 to each $X_k^{(r)}$ individually for a broad range of k. First observe that $\mathbb{E}(X_k^{(r)})$, already computed in Claim 2.1, can be alternatively expressed as

$$\mathbb{E}(X_k^{(r)}) = k - \frac{r\binom{k}{r}}{\binom{rn-1}{r-1}} = k - \frac{k\binom{k-1}{r-1}}{\binom{rn-1}{r-1}} = k\left(1 - \frac{\binom{k-1}{r-1}}{\binom{rn-1}{r-1}}\right)$$

$$= k\left(1 - \frac{k-1}{rn-1}\cdot\frac{k-2}{rn-2}\cdot\ldots\cdot\frac{k-r+1}{rn-r+1}\right)$$

and thus

$$k\left(1 - \left(\frac{k}{rn}\right)^{r-1}\right) \leqslant \mathbb{E}(X_k^{(r)}) \leqslant k\left(1 - \left(\frac{k-r}{rn-r}\right)^{r-1}\right). \tag{1}$$

Also, note that

$$\left(\frac{k-r}{rn-r}\right)^{r-1} \geqslant \left(\frac{k-r}{rn}\right)^{r-1} = \left(\frac{k}{rn} - \frac{1}{n}\right)^{r-1} = \left(\frac{k}{rn}\right)^{r-1} - O\left(\frac{k^{r-2}}{n^{r-1}}\right) = \left(\frac{k}{rn}\right)^{r-1} - O\left(\frac{1}{k}\right),$$

as $k \leqslant rn$. Hence,

$$\mathbb{E}(X_k^{(r)}) = k\left(1 - \left(\frac{k}{rn}\right)^{r-1}\right) + O(1).$$

Since $f(x) = x - \frac{x^r}{(rn)^{r-1}}$ for $1 \leqslant x \leqslant rn$ achieves its maximum at $x_0 = r^{\frac{r-2}{r-1}}n$ and $f(x_0) = \frac{(r-1)n}{r^{1/(r-1)}}$, it follows that for each k,

$$\mathbb{E}(X_k^{(r)}) \leqslant f(k_0) = \frac{(r-1)n}{r^{1/(r-1)}} + O(1). \tag{2}$$

where $k_0 = \lceil r^{\frac{r-2}{r-1}}n \rceil$. Moreover, by the L-H-S of (1),

$$\mathbb{E}(X_{k_0}^{(r)}) \geqslant k_0\left(1 - \left(\frac{k_0}{rn}\right)^{r-1}\right) \geqslant r^{\frac{r-2}{r-1}}n\left(1 - \left(\frac{r^{\frac{r-2}{r-1}}n + 1}{rn}\right)^{r-1}\right)$$

$$= r^{\frac{r-2}{r-1}} n \left(1 - \left(\frac{1}{r^{1/(r-1)}} + \frac{1}{rn} \right)^{r-1} \right) = \frac{(r-1)n}{r^{1/(r-1)}} + O(1).$$

Let

$$n^{4/5} \leqslant k \leqslant rn - n^{4/5}. \tag{3}$$

Observe that for any k outside the above range,

$$X_k^{(r)} \leqslant \min\{k, (rn-k)(r-1)\} \leqslant (r-1)n^{4/5}$$

We next show that for each k satisfying (3) the random variable $X_k^{(r)}$ is highly concentrated around its mean. Indeed, for k in this range, by the L-H-S of (1),

$$\mathbb{E}X_k^{(r)} = k - \frac{r\binom{k}{r}}{\binom{rn-1}{r-1}} \geqslant k \left(1 - \frac{k}{rn} \right) \geqslant n^{4/5} \left(1 - \frac{rn - n^{4/5}}{rn} \right) = \frac{n^{3/5}}{r}.$$

Thus, Lemma 2.1, applied to $h(\pi) = x_k(M_\pi)$ with $N = rn$, $c = 2(r-1)$, and $\eta = \frac{1}{2}C\sqrt{n \log n}$ implies that for each k satisfying (3)

$$\mathbb{P}\left(|X_k^{(r)} - \mathbb{E}(X_k^{(r)})| \geqslant \frac{1}{2}C\sqrt{n \log n} \right) \leqslant 2\exp\left\{ -\frac{C^2 \log n}{2c^2 r} \right\} = o(n^{-1}), \tag{4}$$

provided $C > 2(r-1)\sqrt{2r}$.

Let us set $R = \frac{(r-1)}{r^{1/(r-1)}}$ and keep notation $\eta = \frac{1}{2}C\sqrt{n \log n}$ for convenience. Recalling that $\mathbb{E}(X_{k_0}^{(r)}) \geqslant Rn + O(1)$, it follows that a.a.s.

$$Y^{(r)} = \max_k X_k^{(r)} \geqslant X_{k_0}^{(r)} \geqslant \mathbb{E}X_{k_0}^{(r)} - \eta \geqslant Rn - 2\eta.$$

On the other hand, by (4), (2), and the union bound,

$$\mathbb{P}\left(Y > Rn + 2\eta \right) \leqslant \mathbb{P}\left(\exists k : X_k^{(r)} > Rn + 2\eta \right) \leqslant \sum_k \mathbb{P}\left(|X_k^{(r)} - \mathbb{E}(X_k^{(r)})| \geqslant \eta \right) = o(1),$$

where the last sum is over all k in the range (3). This completes the proof of Theorem 1.4. $\qquad\square$

3. Enumeration of matchings with a given sock number

Let $\mathcal{M}(n)$ be the set of all matchings of $[2n]$. For $M \in \mathcal{M}(n)$ define the *sock number* of M, $y(M) := \max_k x_k(M)$, to be the size of the largest bipartite sub-matching of M (a particular instance of the random variable Y defined earlier). For $1 \leqslant j \leqslant n$, define

$$\mathcal{M}_{\leqslant j}(n) = \{M \in \mathcal{M}(n) : y(M) \leqslant j\},$$

$$s_{\leqslant j}(n) = |\mathcal{M}_{\leqslant j}(n)| \quad \text{and} \quad s_j(n) = s_{\leqslant j}(n) - s_{\leqslant j-1}(n),$$

where we assume that $s_{\leqslant 0}(n) = 0$. Thus, $s_j(n)$ counts the number of matchings of size n with sock number exactly j.

It seems to be a very challenging problem to determine the values of $s_j(n)$. Here we just take the first step in this direction by finding numbers $s_j(n)$ for $j \leqslant 2$ and $j \geqslant n-1$. As a technical tool we introduce the *Dyck trace* of M, that is a binary sequence $tr(M) = t_1 \dots t_{2n}$, $t_i \in \{1, -1\}$, where $t_1 = 1$ if i is the left end of an edge of M and $t_i = -1$ if i is the right end of an edge of M. For instance, if $M = ABCCDBDA$, then $tr(M) = (1, 1, 1, -1, 1, -1, -1, -1)$. The name comes from *Dyck sequences* which are binary sequences, with say n 1's and n (−1)'s, where each prefix has at least as many 1's as (−1)'s, equivalently, for each $1 \leqslant k \leqslant 2n$, we have $\sum_{i=1}^k t_i \geqslant 0$. Dyck sequences are enumerated by the Catalan numbers $\frac{1}{n+1}\binom{2n}{n}$.

Clearly, several matchings may have the same Dyck trace. More precisely, define a *run* in a sequence as a maximal block of identical elements. If the consecutive runs of 1's and -1's in a Dyck sequence have lengths $l_1, r_1, l_2, r_2, \dots$, then there are exactly $(l_1)_{r_1} \cdot (l_1 + l_2 - r_1)_{r_2} \cdots$ matchings with this Dyck trace, where $(l)_r = l(l-1) \cdots (l-r+1)$ is the falling factorial. (In the above example, $l_1 = 3, r_1 = 1, l_2 = 1, r_2 = 3$, so there are $(3)_1 \cdot (3)_3 = 3 \cdot 3 \cdot 2 \cdot 1 = 18$ matchings with the same Dyck trace as M.) Among them, there is exactly one crossing-free and exactly one nesting-free matching which, by the way, proves that both these subfamilies of ordered matchings are enumerated by the Catalan numbers (see [13]).

Let us note that the sock number of M, $y(M)$, equals the *height* of its trace $tr(M)$, that is, $\max_k \sum_{i=1}^k t_i$. Dyck sequences with height at most j have been enumerated by a recurrence relation (see https://oeis.org/A080934) but no closed formula was found. It seems that determining $s_j(n)$ is not an easier task.

Remark 3.1. As mentioned above, ordered matchings M of $[2n]$ with no nestings are uniquely determined by their Dyck traces $tr(M)$, and the same is true for matchings with no crossings. Thus, the random matching $\mathbb{RM}(n)$ conditioned on containing no nestings, as well as the one conditioned on containing no crossings, is equivalent to a random Dyck sequence $\mathbb{RD}(n)$ of length $2n$. Thus, in these conditional spaces, the random variable $y(\mathbb{RM}(n))$ has the same distribution as the height $h(\mathbb{RD}(n))$. As for the latter, it can be shown by a standard application of Chernoff's inequality (see Ineq. (2.9) together with Theorem 2.10 in [9]) that a.a.s., $h(\mathbb{RD}(n)) = O(\sqrt{n \log n})$, a significant drop from $\Theta(n)$ established in Theorem 1.2 for unconditional $\mathbb{RM}(n)$.

Trivially, $s_1(n) = 1$, as $M_1 = A_1 A_1 A_2 A_2 \cdots A_n A_n$ is the only matching with $y(M) = 1$, and $s_n(n) = n!$, as $y(M) = n$ if and only if M is bipartite, and so M is determined by the order of the right endpoints of its edges. Our two final results provide formulae for $s_2(n)$ and $s_{n-1}(n)$, respectively.

Proposition 3.1. *We have $s_{\leqslant 2}(1) = 1$ and, for each $n \geqslant 2$,*
$$s_{\leqslant 2}(n) = 3s_{\leqslant 2}(n-1).$$

Consequently, $s_{\leqslant 2}(n) = 3^{n-1}$ and $s_2(n) = s_{\leqslant 2}(n) - s_1(n) = 3^{n-1} - 1$.

Proof. Note that for every $M \in \mathcal{M}_{\leqslant 2}(n)$ other than M_1, the longest run in $tr(M)$ has length two. Let $\mathcal{M}_{\leqslant 2}(n, t)$ be the set of all matchings in $M \in \mathcal{M}_{\leqslant 2}(n)$ for which in $tr(M)$ the first (from the left) 11-run occurs at positions $2t - 1, 2t$. (Note that any such run has to begin at an odd position.) The digit -1 which must follow this 11 in $tr(M)$ may represent the right end of an edge whose left end is either $2t - 1$ or $2t$. Once this is decided, we are looking at a fresh sub-matching of size $n - t$. Thus,
$$s_{\leqslant 2}(n) = |\mathcal{M}_{\leqslant 2}(n)| = 1 + \sum_{t=1}^{n} |\mathcal{M}_{\leqslant 2}(n, t)| = 1 + 2\sum_{t=1}^{n} |\mathcal{M}_{\leqslant 2}(n - t)|,$$
and the stated recurrence follows by canceling all common terms in $s_{\leqslant 2}(n) - s_{\leqslant 2}(n-1)$. $\square$

Proposition 3.2. *For each $n \geqslant 1$,*
$$s_{n-1}(n) = (n-1)^2 (n-1)!.$$

Proof. We have $y(M) = n - 1$ if and only if there is exactly one edge $e \in M$ with both ends in $[n]$, that is, $e \subset [n]$. If $e \subset [n-1]$, then the $n-1$ elements of $[n+1] \setminus e$ can be matched arbitrarily with those in $[2n] \setminus [n+1]$. So, there are $\binom{n-1}{2} \times (n-1)!$ such matchings M. If $e = \{i, n\}$ for some $i \in [n-1]$, then the $n-2$ vertices in $[n-1] \setminus \{i\}$ can be matched arbitrarily with some $n-1$ elements of $[2n] \setminus [n]$, leaving out two elements to form the final edge of M. So, in this case there are $(n-1) \times (n)_{n-2}$ such matchings M. Summing up the two quantities yields the required formula. $\square$

Problem 3.1. *Determine $s_k(n)$ (by finding a compact formula, a recurrence, or a generating function) for all values of n and k.*

It seems that even the exact determination of $s_3(n)$ may be quite challenging.

4. Final remarks

We conclude the paper with a number of far-reaching generalizations of Bosek's socks problem. We begin with a couple of problems that lack the sequential aspect of Bosek's questions but can be viewed as their Turán-type counterparts. Recall the dual nature of the random variable Y defined in the Introduction which on the one hand is the maximum value of a sockuence, while on the other hand, it is the size of the largest bipartite sub-matching, equivalently, the largest sub-matching not containing pattern $AABB$.

Two words over a finite alphabet are deemed to be *isomorphic* if they are the same up to a permutation of the letters. For instance, the words $ABBA$ and $DAAD$ are isomorphic. (Note that for r-fold Gauss words this is equivalent to the order-preserving isomorphism of the corresponding ordered r-matchings.)

Let $\mathcal{F}$ be a fixed set of forbidden ordered r-matchings (or its corresponding Gauss words). An ordered r-matching M is $\mathcal{F}$-*free* if M does not contain a sub-matching (order) isomorphic to F for any $F \in \mathcal{F}$. What is the maximum size of an $\mathcal{F}$-free sub-matching of $\mathbb{RM}^{(r)}(n)$? As mentioned above, for $r = 2$, Bosek's original question corresponds to the case of $\mathcal{F} = \{AABB\}$. Let us denote by $ex_{\mathcal{F}}(M)$ the maximum number of edges in an $\mathcal{F}$-free sub-matching of M.

Problem 4.1. *For $r \geqslant 2$, let $\mathcal{F}$ be a set of forbidden ordered r-matchings. Estimate the expected value of $ex_{\mathcal{F}}(\mathbb{RM}^{(r)}(n))$.*

Next, consider a special case of the above problem in which all elements of $\mathcal{F}$ are patterns, that is, pairs of edges. Given $r \geqslant 2$, there are exactly $\frac{1}{2}\binom{2r}{r}$ ways, called r-*patterns*, in which a pair of disjoint edges of order r may intertwine as ordered vertex sets. In this restricted setting, it is equivalent, and perhaps more natural, to emphasize not the forbidden patterns but the complementary set of those that are allowed to be present. For a set of r-patterns $\mathcal{P}$, a $\mathcal{P}$-*clique* is defined as an r-matching whose all pairs of edges form patterns P belonging to $\mathcal{P}$. Let us denote by $z_{\mathcal{P}}(M)$ the maximum number of edges in a $\mathcal{P}$-clique contained in M.

Problem 4.2. *For $r \geqslant 2$, let $\mathcal{P}$ be a set of r-patterns. Estimate the expected value of $z_{\mathcal{P}}(\mathbb{RM}^{(r)}(n))$.*

Recall that, for $r = 2$, a bipartite matching is one without alignments $AABB$, and so the above problem in that case, that is, when $\mathcal{P} = \{ABAB, ABBA\}$ is solved by Theorem 1.2. On the other hand, it can be easily shown by the first-moment method that if one instead forbids nestings or crossings, that is, if $\mathcal{P} = \{AABB, ABBA\}$ or $\mathcal{P} = \{AABB, ABAB\}$, then the answer is just $O(\sqrt{n})$ (and not $\Theta(n)$) – the same when $|\mathcal{P}| = 1$ (see [6]).

For larger r, we know a typical order of magnitude of $z_{\mathcal{P}}(\mathbb{RM}^{(r)}(n))$ when $\mathcal{P} = \{P\}$. (In such case we use the simplified notation $z_P(\mathbb{RM}^{(r)}(n))$.) To state the result, we have to distinguish a special subfamily of patterns. Call an r-pattern P *collectable* if for each $k \geqslant 1$ there exists a $\{P\}$-clique of size k. For $r \geqslant 3$, not every r-pattern is collectable. For instance, the 3-pattern $P_0 = AABABB$ is the unique not collectable 3-pattern (it fails already at $k = 3$). In fact, as shown in [4], there are exactly 3^{r-1} collectable r-patterns.

By definition, for every non-collectable pattern P, we have $z_P(\mathbb{RM}^{(r)}(n)) = O(1)$. On the other hand, it was proved in [4] that for every collectable r-pattern P a.a.s. $z_P(\mathbb{RM}^{(r)}(n)) = \Theta(n^{1/r})$. (In [1] the constants in front of $n^{1/r}$ have been, in principle, determined.)

Let us further restrict to r-partite patterns and sub-matchings (i.e., those whose interval chromatic number is r). There are exactly 2^{r-1} r-partite r-patterns and they are all collectable. Let $\mathcal{Q}^{(r)}$ be the family of all of them. In particular,

$$\mathcal{Q}^{(3)} = \{ABABAB, ABBAAB, ABABBA, ABBABA\}.$$

Problem 4.3. *Determine a.a.s. the largest size of an r-partite sub-matching of the random r-matching $\mathbb{RM}^{(r)}(n)$, that is, determine $z_{\mathcal{Q}^{(r)}}(\mathbb{RM}^{(r)}(n))$. Further, determine a.a.s. the order of magnitude of $z_{\mathcal{P}}(\mathbb{RM}^{(r)}(n))$, when $\mathcal{P} \subset \mathcal{Q}^{(r)}$ is a subset of the set of r-partite patterns.*

We believe that $z_{\mathcal{Q}^{(r)}}(\mathbb{RM}^{(r)}(n)) = \Theta(n)$ but would like to pinpoint the multiplicative constant – as in Theorem 1.4. For $r = 3$ we further conjecture that whenever $\mathcal{P} \subset \mathcal{Q}^{(3)}$ and $|\mathcal{P}| = 2$, a.a.s. $z_{\mathcal{P}}(\mathbb{RM}^{(3)}(n)) = \Theta(n^{1/2})$.

Finally, let us pose a very general problem that reflect the sequential aspect of Bosek's question. Here we consider general words over finite alphabets, not just Gauss words. Suppose that $\mathcal{J}$ is a finite family of words (to be *persecuted*). For a given word U, let $U/\mathcal{J}$ be the (scattered) sub-word of U obtained by deleting successively (from left to right) one isomorphic copy of an element $J \in \mathcal{J}$ at a time. More precisely, we scan the letters one by one and once we see such a copy emerging, we remove all letters forming it. If more than one copy of the elements in $\mathcal{J}$ has been revealed at that moment, we delete, say, the one that is farther to the left (destroying, in fact, all of them).

For instance, if $\mathcal{J} = \{ABA\}$ and $U = ABCADBCDA$, then $U/\mathcal{J} = BDA$. Indeed, the first removal (of ABA, and not of ACA) takes place after four letters have been scanned and leaves the sub-word $CDBCDA$ which we continue to scan from the left; the second removal (of CDC, and not of CBC) leaves the final sub-word BDA.

Let $y_{\mathcal{J}}(U) = \max_i |U_i/\mathcal{J}|$, where U_i is the *prefix* of U of length i. (In the above example, $y_{\mathcal{J}}(U) = 3$.) Notice that when U is a Gauss word and $\mathcal{J} = \{AA\}$, the number $y_{\mathcal{J}}(U)$ coincides with the sock number $y(M)$ of the corresponding ordered matching M.

Let $\mathbb{RW}_k(n)$ be any model of a random word of length n over an alphabet of size k. (E.g., in [5], two natural uniform models are defined).

Problem 4.4. *For $k \geqslant 2$ and a fixed persecuted family $\mathcal{J}$, estimate the expected value of $y_{\mathcal{J}}(\mathbb{RW}_k(n))$.*

Note that Bosek's problem for r-sets of socks asks for $y_{\mathcal{J}}(\mathbb{RW}_k(n))$, where $\mathcal{J} = \{A^r\}$, $k = n$, $n := rn$, and $\mathbb{RW}_k(n)$ is a random r-fold Gauss word of length rn.

Acknowledgements

References

[1] M. Anastos, Z. Jin, M. Kwan, and B. Sudakov, *Extremal, enumerative and probabilistic results on ordered hypergraph matchings*, arXiv:2308.12268.

[2] J. Baik and E. M. Rains, *The asymptotics of monotone subsequences of involutions*, Duke Math. J. 109:2 (2001), 205–281.

[3] B. Bosek, Personal communication.

[4] A. Dudek, J. Grytczuk, and A. Ruciński, *Erdős-Szekeres type theorems for ordered uniform matchings*, arXiv:2301.02936.

[5] A. Dudek, J. Grytczuk, and A. Ruciński, *Long twins in random words*, Ann. Comb. 27:3 (2023), 749–768.

[6] A. Dudek, J. Grytczuk, and A. Ruciński, *Ordered unavoidable sub-structures in matchings and random matchings*, Electron. J. Combin. 31:2 (2024), Paper No. 2.15.

[7] A. Frieze and B. Pittel, *Perfect matchings in random graphs with prescribed minimal degree*, Mathematics and computer science. III, Birkhäuser, Basel, 2004, 95–132.

[8] Z. Füredi, T. Jiang, A. Kostochka, D. Mubayi, and J. Verstraëte, *Extremal problems for convex geometric hypergraphs and ordered hypergraphs*, Canad. J. Math. 73:6 (2021), 1648–1666.

[9] S. Janson, T. Łuczak, and A. Ruciński, *Random graphs*, Wiley-Interscience Series in Discrete Mathematics and Optimization, Wiley-Interscience, New York, 2000.

[10] J. Justicz, E. R. Scheinerman, and P. M. Winkler, *Random intervals*, Amer. Math. Monthly 97:10 (1990), 881–889.

[11] C. McDiarmid, *Probabilistic methods for algorithmic discrete mathematics* (M. Habib, C. McDiarmid, J. Ramirez-Alfonsin, and B. Reed, eds.), Algorithms and Combinatorics 16, Springer-Verlag, Berlin, 1998.

[12] E. R. Scheinerman, *Random interval graphs*, Combinatorica 8:4 (1988), 357–371.

[13] R. P. Stanley, *Catalan numbers*, Cambridge University Press, New York, 2015.