Discrete Applied Mathematics 210 (2016) 75–87

Contents lists available at ScienceDirect

Discrete Applied Mathematics

journal homepage: www.elsevier.com/locate/dam

Repetition-free longest common subsequence of random sequences

Cristina G. Fernandes [a], Marcos Kiwi [b,*]

[a] Computer Science Department, Universidade de São Paulo, Brazil

[b] Depto. Ingeniería Matemática & Ctr. Modelamiento Matemático UMI 2807, U. Chile, Chile

ARTICLE INFO

Article history:
Received 9 December 2013
Received in revised form 2 July 2015
Accepted 6 July 2015
Available online 6 August 2015

Keywords:
Repetition-free subsequence
Common subsequence
Random sequences

ABSTRACT

A repetition-free Longest Common Subsequence (LCS) of two sequences x and y is an LCS of x and y where each symbol may appear at most once. Let R denote the length of a repetition-free LCS of two sequences of n symbols each one chosen randomly, uniformly, and independently over a k-ary alphabet. We study the asymptotic, in n and k, behavior of R and establish that there are three distinct regimes, depending on the relative speed of growth of n and k. For each regime we establish the limiting behavior of R. In fact, we do more, since we actually establish tail bounds for large deviations of R from its limiting behavior.

Our study is motivated by the so called exemplar model proposed by Sankoff (1999) and the related similarity measure introduced by Adi et al. (2010). A natural question that arises in this context, which as we show is related to long standing open problems in the area of probabilistic combinatorics, is to understand the asymptotic, in n and k, behavior of parameter R.

1. Introduction

Several of the genome similarity measures considered in the literature either assume that the genomes do not contain gene duplicates, or work efficiently only under this assumption. However, several known genomes do contain a significant amount of duplicates. (See the review on gene and genome duplication by Sankoff [19] for specific information and references.) One can find in the literature proposals to address this issue. Some of these proposals suggest to filter the genomes, throwing away part or all of the duplicates, and then applying the desired similarity measure to the filtered genomes. (See [2] for a description of different similarity measures and filtering models for addressing duplicates.)

Sankoff [18], trying to take into account gene duplication in genome rearrangement, proposed the so called exemplar model, which is one of the filtering schemes mentioned above. In this model, one searches, for each family of duplicated genes, an exemplar representative in each genome. Once the representative genes are selected, the other genes are disregarded, and the part of the genomes with only the representative genes is submitted to the similarity measure. In this case, the filtered genomes do not contain duplicates, therefore several of the similarity measures (efficiently) apply. Of course, the selection of the exemplar representative of each gene family might affect the result of the similarity measure. Following the parsimony principle, one wishes to select the representatives in such a way that the resulting similarity is as good as possible. Therefore, each similarity measure induces an optimization problem: how to select exemplar representatives of each gene family that result in the best similarity according to that specific measure.

* Corresponding author.
E-mail addresses: cris@ime.usp.br (C.G. Fernandes), mkiwi@dim.uchile.cl (M. Kiwi).

http://dx.doi.org/10.1016/j.dam.2015.07.005
0166-218X/

The length of a Longest Common Subsequence (LCS) is a well-known measure of similarity between sequences. In particular, in genomics, the length of an LCS is directly related to the so called edit distance between two sequences when only insertions and deletions are allowed, but no substitution. This similarity measure can be computed efficiently in the presence of duplicates (the classical dynamic programming solution to the LCS problem takes quadratic time, however, improved algorithms are known, specially when additional complexity parameters are taken into account — for a comprehensive comparison of well-known algorithms for the LCS problem, see [4]). Inspired by the exemplar model above, some variants of the LCS similarity measure have been proposed in the literature. One of them, the so called *exemplar* LCS [6], uses the concept of mandatory and optional symbols, and searches for an LCS containing all mandatory symbols. A second one is the so called *repetition-free* LCS [1], that requires each symbol to appear at most once in the subsequence. Some other extensions of these two measures were considered under the name of *constrained* LCS and *doubly-constrained* LCS [7]. All of these variants were shown to be hard to compute [1,5–7], so some heuristics and approximation algorithms for them were proposed and experimentally tested [1,6,14,10].

Specifically, the notion of repetition-free LCS was formalized by Adi et al. [1] as follows. They consider finite sets, called *alphabets*, whose elements are referred to as *symbols*, and then they define the RFLCS problem as: Given two sequences x and y, find a repetition-free LCS of x and y. We write RFLCS (x, y) to refer to the RFLCS problem for a generic instance consisting of a pair (x, y), and we denote by Opt(RFLCS(x, y)) the length of an optimal solution of RFLCS (x, y). In their paper, Adi et al. showed that RFLCS is MAX SNP-hard, proposed three approximation algorithms for RFLCS, and presented an experimental evaluation of their proposed algorithms, using for the sake of comparison an exact (computationally expensive) algorithm for RFLCS based on an integer linear programming formulation of the problem.

Whenever a problem such as the RFLCS is considered, a very natural question arises: What is the expected value of Opt(RFLCS(x, y))? (where expectation is taken over the appropriate distribution over the instances (x, y) one is interested in). It is often the case that one has little knowledge of the distribution of problem instances, except maybe for the size of the instances. Thus, an even more basic and often relevant issue is to determine the expected value taken by Opt(RFLCS(x, y)) for uniformly distributed choices of x and y over all strings of a given length over some fixed size alphabet (say each sequence has n symbols randomly, uniformly, and independently chosen over a k-ary alphabet Σ). Knowledge of such an average case behavior is a first step in the understanding of whether a specific value of Opt(RFLCS(x, y)) is of relevance or could be simply explained by random noise. The determination of this latter average case behavior in the asymptotic regime (when the length n of the sequences x and y go to infinity) is the main problem we undertake in this article. Specifically, let $R_n = R_n(x, y)$ denote the length of a repetition-free LCS of two sequences x and y of n symbols randomly, uniformly, and independently chosen over a k-ary alphabet. Note that the random variable R_n is simply the value of Opt(RFLCS(x, y)). We are interested in determining (approximately) the value of $\mathbb{E}\left(R_n\right)$ as a function of n and k, for very large values of n.

One of the results established in this article is that the behavior of $\mathbb{E}\left(R_n\right)$ depends on the way in which n and k are related. In fact, if k is fixed, it is easy to see that $\mathbb{E}\left(R_n\right)$ tends to k when n goes to infinity (simply because any fix permutation of a k-ary alphabet will appear in a sufficiently large sequence of uniformly and independently chosen symbols from the alphabet). Thus, the interesting cases arise when $k = k(n)$ tends to infinity with n. However, the speed at which $k(n)$ goes to infinity is of crucial relevance in the study of the behavior of $\mathbb{E}\left(R_n\right)$. We identify three distinct growth regimes depending on the asymptotic dependency between n and $k\sqrt{k}$. Specifically, we establish the next result[1]:

Theorem 1. *The following holds:*

- *If* $n = \omega(\sqrt{k})$ *and* $n = o(k\sqrt{k})$, *then* $\lim_{n\to\infty} \frac{\mathbb{E}\left(R_n\right)}{n/\sqrt{k(n)}} = 2$.
- *If* $n = \frac{1}{2}\rho k\sqrt{k}$ *for* $\rho > 0$, *then* $\liminf_{n\to\infty} \frac{\mathbb{E}\left(R_n\right)}{k(n)} \geq 1 - e^{-\rho}$. *(By definition* $R_n \leq k(n)$.*)*

  Moreover, if $n = \omega(k\sqrt{k})$, *then* $\lim_{n\to\infty} \frac{\mathbb{E}\left(R_n\right)}{k(n)} = 1$.

The main results of this article are obtained by relating the asymptotic average case behavior of $\mathbb{E}\left(R_n\right)$ with that of the length $L_n = L_n(x, y)$ of a Longest Common Subsequence (LCS) of two sequences x and y of n symbols chosen randomly, uniformly, and independently over a k-ary alphabet. A simple (well-known) fact concerning L_n is that $\mathbb{E}\left(L_n\right)/n$ tends to a constant, say γ_k, when n goes to infinity. The constant γ_k is known as the Chvátal–Sankoff constant. A long standing open problem is to determine the exact value of γ_k for any fixed $k \geq 2$. However, Kiwi, Loebl, and Matoušek [17] proved that $\gamma_k\sqrt{k} \to 2$ as $k \to \infty$ (which positively settled a conjecture due to Sankoff and Mainville [20]).

We now give an informal and intuitive justification for each of the claims stated in Theorem 1. As pointed out above, in [17], it was shown that, under some conditions on the speed of growth of $k = k(n)$, the expected length of an LCS of two length n sequences randomly, uniformly, and independently chosen over a k-ary alphabet, is roughly $2n/\sqrt{k}$. When $n = \omega(\sqrt{k}) \cap o(k\sqrt{k})$, we see that $2n/\sqrt{k} = \omega(1) \cap o(k)$. If the k-ary symbols that belong to an LCS show up more or

[1] Adhering to standard notation, for functions f and g defined over the non-negative integers, g always non-zero, we say that $f(n) = \omega(g(n))$ if $|f(n)|/|g(n)|$ tends to infinity when $n \to \infty$.

less with the same frequency, then one expects that there will be few repetitions in an LCS of size $o(k)$, and thus expects a repetition-free LCS of size $2n/\sqrt{k}$. In contrast, when $n = \frac{1}{2}\rho k\sqrt{k}$, $\rho > 0$, we have that $2n/\sqrt{k} = \rho k$. Under the same foregoing assumption concerning the frequency of occurrence of symbols in an LCS, we should expect some repetition of symbols. If one keeps one of each symbol occurrence of the roughly ρk length LCS, one would be left with a repetition-free LCS. A coupon collector type argument suggests that the length of such repetition-free LCS will be at least $(1 - e^{-\rho})k$. Finally, it is natural for the remaining $n = \omega(k\sqrt{k})$ case to be well approximated by the $n = \frac{1}{2}\rho k\sqrt{k}$ case for very large ρ, thus suggesting that an LCS will end up being of size k.

It is worth mentioning that the above discussion refers to sequences randomly, uniformly, and independently chosen over an alphabet. Genome sequences might not be characterized well by those, as duplications are evolutionary events, and occur in several different levels, such as duplication of single genes or of an entire chromosome. See [22, Chaps. 7 and 8] for a more complete discussion on this.

Coming back to the description of this article's main contributions, for each of the three different regimes of Theorem 1, we also establish so called large deviation bounds which capture how unlikely it is for R_n to deviate too much from its expected value. Specifically, we establish the following:

Theorem 2. *The following holds:*

- *If* $n = \omega(\sqrt{k})$ *and* $n = o(k\sqrt{k})$, *then for every* $0 < \xi \le 1$ *there is a constant* $k_0 = k_0(\xi)$ *such that, for all* $k > k_0$,
$$\mathbb{P}\left(R_n \le (1-\xi)2n/\sqrt{k}\right) \le 2e^{-\frac{1}{10}\xi^2 2n/\sqrt{k}}.$$
- *Let* $\rho > 0$ *and* $0 < \xi < 1$. *If* $n = \frac{1}{2}\rho k\sqrt{k}$, *then there is a constant* $k_0 = k_0(\rho, \xi)$ *such that, for all* $k > k_0$,
$$\mathbb{P}\left(R_n \le (1-\xi)k(1-e^{-\rho})\right) \le 2e^{-\frac{\xi^2}{32(1+\xi/12)}k(1-e^{-\rho})} \le 2e^{-\frac{1}{35}\xi^2 k(1-e^{-\rho})}.$$
- *If* $n = (\frac{1}{2} + \xi)k\sqrt{k}\ln k$ *for some* $\xi > 0$, *then there is a constant* $k_0 = k_0(\xi)$ *such that, for all* $k > k_0$,
$$\mathbb{P}\left(R_n \ne k\right) \le \frac{2}{k^{\xi}}.$$

The preceding stated result is stronger than the one derived by standard concentration type inequalities. Indeed, an application of Azuma's inequality [21, Theorem 1.3.1] in exactly the same way as applied to obtain large deviation bounds for the length of an LCS of two randomly and uniformly chosen equal length sequences (see for example [21, Section 1.3]) yields, for $t > 0$,

$$\mathbb{P}\left(\left|R_n - \mathbb{E}\left(R_n\right)\right| \ge t\right) \le 2e^{-t^2/8n}.$$

By Theorem 1 we then get that the probability that $R_n \le (1-\xi)\mathbb{E}\left(R_n\right)$ is at most $e^{-O(1)\xi^2 n/k}$ when $n = \omega(\sqrt{k}) \cap o(k\sqrt{k})$, and at most $e^{-O(1)\xi^2\sqrt{k}(1-e^{-\rho})^2/\rho}$ when $n = \frac{1}{2}\rho k\sqrt{k}$. In both cases, Theorem 2 gives a sharper lower tail deviation bound.

Throughout this article we build upon [17], and draw connections with another intensively studied problem concerning Longest Increasing Subsequences (LIS) of randomly chosen permutations (also known as Ulam's problem). Probably even more significant is the fact that our analysis partly elicits a typical structure of one of the large repetition-free common subsequences of two length n sequences randomly, uniformly, and independently chosen over a k-ary alphabet.

In order to formalize some aspects of our preceding discussion and rigorously state and derive our claims, we first need to introduce terminology, some background material, and establish some basic facts. We start by describing the road-map followed throughout this manuscript.

Organization: This article is organized as follows. In Section 2, we review some classical probabilistic so called urn models and, for the sake of completeness, summarize some of their known basic properties, as well as establish a few others. As our results build upon those of Kiwi, Loebl, and Matoušek [17], we review them in Section 3, and also take the opportunity to introduce some relevant terminology. In Section 4, we formalize the notion of "canonical" repetition-free LCS and show that conditioning on its size, the distribution of the set of its symbols is uniform (among all appropriate size subsets of symbols). Although simple to establish, this result is key to our approach since it allows us to relate the probabilistic analysis of the length of repetition-free LCSs to one concerning urn models. Finally, in Section 5, we establish large deviation type bounds from which Theorem 1 easily follows.

2. Background on urn models

The probabilistic study of repetition-free LCSs we will undertake will rely on the understanding of random phenomena that arises in so called urn models. In these models, there is a collection of urns where balls are randomly placed. Different ways of distributing the balls in the urns, as well as considerations about the (in)distinguishability of urns/balls, give rise to distinct models, often referred to in the literature as occupancy problems (for a classical treatment see [13]). In this section, we describe those urn models we will later encounter, associate to them parameters of interest, and state some basic results concerning their probabilistic behavior.

Henceforth, let k and s be positive integers, and $\vec{s} = (s_1, \ldots, s_b)$ denote a b-dimensional nonnegative integer vector whose coordinates sum up to s, i.e., $\sum_{i=1}^{b} s_i = s$. For a positive integer m, we denote the set $\{1, \ldots, m\}$ by $[m]$.

Consider the following two processes where s indistinguishable balls are randomly distributed among k distinguishable urns.

- **Grouped Urn** $(k, \vec{s})$**-model:** Randomly distribute s balls over k urns, placing a ball in urn j if $j \in S_i$, where $S_1, \ldots, S_b \subseteq [k]$ are chosen randomly and independently so that S_i is uniformly distributed among all subsets of $[k]$ of size s_i.
- **Classical Urn** (k, s)**-model:** Randomly distribute s balls over k urns, so that the ith ball, $i \in [k]$, is placed in an urn uniformly chosen among the k urns, and independently of where the other balls are placed.[2]

Henceforth, let $X^{(k,\vec{s})}$ be the number of empty urns left when the Grouped Urn $(k, \vec{s})$-process ends. Furthermore, let $X_j^{(k,\vec{s})}$ be the indicator of the event that the jth urn ends up empty. Obviously, $X^{(k,\vec{s})} = \sum_{j=1}^{k} X_j^{(k,\vec{s})}$. Similarly, define $Y^{(k,s)}$ and $Y_1^{(k,s)}, \ldots, Y_k^{(k,s)}$ but with respect to the Classical Urn (k, s)-process. Intuitively, one expects that fewer urns will end up empty in the Grouped Urn process in comparison with the Classical Urn process. This intuition is formalized through the following result.

Lemma 3. *Let $\vec{s} = (s_1, \ldots, s_b) \in \mathbb{N}^b$ and $s = \sum_{i=1}^{b} s_i$. Then, the random variable $X^{(k,\vec{s})}$ dominates $Y^{(k,s)}$, i.e., for every $t \geq 0$,*

$$\mathbb{P}\big(X^{(k,\vec{s})} \geq t\big) \leq \mathbb{P}\big(Y^{(k,s)} \geq t\big).$$

Proof. First observe that if $\vec{s} = (1, \ldots, 1) \in \mathbb{N}^s$, then $X^{(k,\vec{s})}$ and $Y^{(k,s)}$ have the same distribution, thence the claimed result trivially holds for such $\vec{s}$. For $\vec{s} = (s_1, \ldots, s_b) \in \mathbb{N}^b$ with $\sum_{i=1}^{b} s_i = s$ and $s_j \geq 2$ for some $j \in [b]$, let $\vec{s'} = (s'_1, \ldots, s'_b, 1) \in \mathbb{N}^{b+1}$ be such that

$$\vec{s'} = (s_1, \ldots, s_{j-1}, s_j - 1, s_{j+1}, \ldots, s_b, 1).$$

Note that $\sum_{i=1}^{b+1} s'_i = s$ and observe that, to establish the claimed result, it will be enough to inductively show that, for every $t \geq 0$,

$$\mathbb{P}\big(X^{(k,\vec{s})} \geq t\big) \leq \mathbb{P}\big(X^{(k,\vec{s'})} \geq t\big). \tag{1}$$

To prove this last inequality, consider the following experiment. Randomly choose $S_1, \ldots, S_b$ as in the Grouped Urn $(k, \vec{s})$-model described above, and distribute s balls in k urns as suggested in the model's description. Recall that $X^{(k,\vec{s})}$ is the number of empty urns left when the process ends. Now, randomly and uniformly choose one of the balls placed in an urn of index in S_j. With probability $\frac{k-(s_j-1)}{k}$, leave it where it is and, with probability $\frac{s_j-1}{k}$, move it to a distinct urn of index in S_j chosen randomly and uniformly. Observe that the number of empty urns cannot decrease. Moreover, note that the experiment just described is equivalent to the Grouped Urn $(k, \vec{s'})$-model, thence the number of empty urns when the process ends is distributed according to $X^{(k,\vec{s'})}$. It follows that (1) holds, thus concluding the proof of the claimed result. □

We will later need upper bounds on the probability that a random variable distributed as $X^{(k,\vec{s})}$ is bounded away (from below) from its expectation, i.e., on so called upper tail bounds for $X^{(k,\vec{s})}$. The relevance of Lemma 3 is that it allows us to concentrate on the rather more manageable random variable $Y^{(k,s)}$, since any upper bound on the probability that $Y^{(k,s)} \geq t$ will also be valid for the probability that $X^{(k,\vec{s})} \geq t$. The behavior of $Y^{(k,s)}$ is a classical thoroughly studied subject. In particular, there are well-known tail bounds that apply to it. A key fact used in the derivation of such tail bounds is that $Y^{(k,s)}$ is the sum of the negatively related 0–1 random variables $Y_1^{(k,s)}, \ldots, Y_k^{(k,s)}$ (for the definition of negatively related random variables, see [15], and the discussion in [15, Example 1]). For convenience of future reference, the next result summarizes the tail bounds that we will use.

Proposition 4. *For all positive integers k and s,*

$$\lambda \stackrel{\text{def}}{=} \mathbb{E}\big(Y^{(k,s)}\big) = k\Big(1 - \frac{1}{k}\Big)^s. \tag{2}$$

Moreover, the following hold:

1. *If $p \stackrel{\text{def}}{=} \lambda/k$ and $q \stackrel{\text{def}}{=} 1 - p$, then for all $a \geq 0$,*

$$\mathbb{P}\big(Y^{(k,s)} \geq \lambda + a\big) \leq \exp\Big(-\frac{a^2}{2(kpq + a/3)}\Big).$$

[2] Note that this model is a particular case of the Grouped Urn model where $b = s$ and $s_1 = \cdots = s_b = 1$.

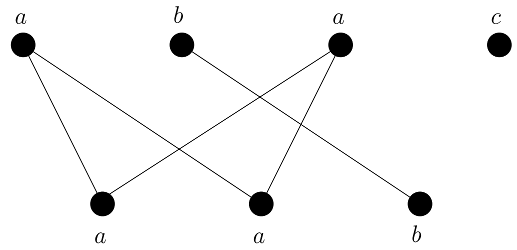

Fig. 1. Graph obtained from $\Sigma(K_{4,3}; 3)$ for the choice of symbols associated (shown close to) each node.

2. *Let* $\xi \geq 0$ *and* $s = (1+\xi)k\ln k$*, then*
$$\mathbb{P}\big(Y^{(k,s)} \neq 0\big) \leq \frac{1}{k^{\xi}}.$$
3. *For all* $a > 0$*,*
$$\mathbb{P}\big(k - Y^{(k,s)} \leq s - a\big) \leq \left(\frac{es^2}{ka}\right)^a.$$

Proof. Since the probability that a ball uniformly distributed over k urns lands in urn j is $1/k$, the probability that none of s balls lands in urn j (equivalently, that $Y^{(k,s)} = 1$) is exactly $(1-1/k)^s$. By linearity of expectation, to establish (2), it suffices to observe that $\mathbb{E}\big(Y^{(k,s)}\big) = \sum_{j=1}^{k} \mathbb{P}\big(Y_j^{(k,s)} = 1\big)$.

Part 1 is just a re-statement of the second bound in (1.4) of [15] taking into account the comments in [15, Example 1].

Part 2 is a folklore result that follows easily from an application of the union bound. For completeness, we sketch the proof. Note that $Y^{(k,s)} \neq 0$ if and only if $Y_j^{(k,s)} \neq 0$ for some $j \in [k]$. Hence, by a union bound and since $1 - x \leq e^{-x}$ for all x,
$$\mathbb{P}\big(Y^{(k,s)} \neq 0\big) \leq \sum_{j\in[k]} \mathbb{P}\big(Y_j^{(k,s)} \neq 0\big) = k\Big(1-\frac{1}{k}\Big)^s \leq ke^{-s/k} = \frac{1}{k^{\xi}}.$$

Finally, let us establish Part 3. Observe that $k - Y^{(k,s)}$ is the number of urns that end up nonempty in the Classical Urn (k, s)-model. Thus, assuming that balls are sequentially thrown, one by one, if $k - Y^{(k,s)} \leq s - a$, then there must be a size a subset $S \subseteq [s]$ of balls that fall in an urn where a previously thrown ball has already landed. The probability that a ball in S ends up in a previously occupied urn, is at most s/k (given that at any moment at most s of the k urns are occupied). So the probability that all balls in S end up in previously occupied urns is at most $(s/k)^a$. Thus, by a union bound, some algebra, and the standard bound on binomial coefficients $\binom{\mu}{\nu} \leq (e\mu/\nu)^{\nu}$,
$$\mathbb{P}\big(k - Y^{(k,s)} \leq s - a\big) \leq \sum_{S\subseteq[s]:|S|=a} \Big(\frac{s}{k}\Big)^a \leq \binom{s}{a}\Big(\frac{s}{k}\Big)^a \leq \left(\frac{es^2}{ka}\right)^a. \quad \square$$

3. Some background on the expected length of an LCS

In [17], pairs of sequences (x, y) are associated to plane embeddings of bipartite graphs, and a common subsequence of x and y to a special class of matching of the associated bipartite graph. Adopting this perspective will also be useful in what follows. In this section, besides reviewing and restating some of the results of [17], we will introduce some of the terminology we shall adhere in what follows.

The *random word model* $\Sigma(K_{r,s}; k)$,[3] as introduced in [17], consists of the following (for an illustration, see Fig. 1): the distribution over the set of subgraphs G of $K_{r,s}$ obtained by uniformly and independently assigning to each vertex of $K_{r,s}$ one of k symbols and keeping in G only those edges whose endpoints are associated to the same symbol. In particular, two nodes u and v belonging to distinct bipartition classes of $K_{r,s}$ end in the same connected component of G if and only if they are assigned the same symbol. That is, G consists of isolated vertices and complete bipartite disjoint subgraphs. Formally, if we let $c(G)$ denote the number of connected components of G, then the probability that G is obtained from $\Sigma(K_{r,s}; k)$ is $k(k-1)\cdots(k-c(G)+1)/k^{r+s}$ if G is a subgraph of $K_{r,s}$ which is a collection of $c(G)$ complete bipartite graphs with no isolated vertices. If G is not a collection of isolated vertices and complete bipartite graphs, then it occurs with probability 0. Otherwise, G is comprised of both isolated vertices and complete bipartite graphs. The probability of its occurrence is given by a more complicated expression which we omit.

Following [17], two distinct edges ab and $a'b'$ of G are said to be *noncrossing* if a and a' are in the same order as b and b'. In other words, if $a < a'$ and $b < b'$, or $a' < a$ and $b' < b$. A matching of G is called *noncrossing* if every distinct pair of its edges is noncrossing.

[3] Remember that $K_{r,s}$ denotes the complete bipartite graph with two bipartition classes, one of size r and the other of size s.

Henceforth, for a bipartite graph G, we denote by $L(G)$ the number of edges in a maximum size (largest) noncrossing matching of G. To G chosen according to $\Sigma(K_{n,n};k)$ we can associate two sequences of length n, say $x(G)$ and $y(G)$, one for each of the bipartition sides of G, consisting of the symbols associated to the vertices of $K_{n,n}$. Note that $x(G)$ and $y(G)$ are uniformly and independently distributed sequences of n symbols over a k-ary alphabet. Observe that, if G is chosen according to $\Sigma(K_{n,n};k)$, then $L(G)$ is precisely the length of an LCS of its two associated sequences $x(G)$ and $y(G)$, and vice versa. Formally, $L(G)=L_n(x(G),y(G))$, where $L_n(\cdot,\cdot)$ is as defined in the introductory section.

Among other things, in [17], it is shown that $L(\Sigma(K_{n,n};k))\sqrt{k}/n$ is approximately equal to 2 when n and k are very large, provided that n is "sufficiently large" compared to k. This result is formalized in the following:

Theorem 5 (*Kiwi, Loebl, and Matoušek [17]*). *For every $\epsilon>0$, there exist k_0 and C such that, for all $k>k_0$ and all n with $n>C\sqrt{k}$,*

$$(1-\epsilon)\cdot\frac{2n}{\sqrt{k}}\le\mathbb{E}\big(L(\Sigma(K_{n,n};k))\big)\le(1+\epsilon)\cdot\frac{2n}{\sqrt{k}}.\tag{3}$$

Moreover, there is an exponentially small tail bound; namely, for every $\epsilon>0$, there exists $c>0$ such that, for k and n as above,

$$\mathbb{P}\left(\left|L(\Sigma(K_{n,n};k))-\frac{2n}{\sqrt{k}}\right|\ge\epsilon\frac{2n}{\sqrt{k}}\right)\le e^{-cn/\sqrt{k}}.$$

Observe now that a graph G chosen according to $\Sigma(K_{n,n};k)$ has symbols, from a k-ary alphabet, implicitly associated to each of its nodes (to the jth node of each side of G the jth symbol of the corresponding sequence $x(G)$ or $y(G)$). Furthermore, the endpoints of an edge e of G must, by construction, be associated to the same symbol, henceforth referred to the symbol associated to e. We say that a noncrossing matching of G is *repetition-free* if the symbols associated to its edges are all distinct, and we denote by $R(G)$ the number of edges in a maximum size (largest) repetition-free noncrossing matching of G. If G is chosen according to $\Sigma(K_{n,n};k)$, then $R(G)$ is precisely the length of a repetition-free LCS of its two associated sequences $x(G)$ and $y(G)$, and vice versa. Formally, $R(G)=R_n(x(G),y(G))$, where again $R_n(\cdot,\cdot)$ is as defined in the introductory section. Summarizing, we have reformulated the repetition-free LCS problem as an equivalent one, but concerning repetition-free noncrossing matchings. This justifies why, from now on, we will speak interchangeably about repetition-free LCSs and repetition-free noncrossing matchings.

Clearly, for every G in the support of $\Sigma(K_{n,n};k)$ we always have that $R(G)\le L(G)$. So, the upper bound in (3) for $\mathbb{E}\big(L(\Sigma(K_{n,n};k))\big)$ and the upper tail bound for $L(\Sigma(K_{n,n};k))$ of Theorem 5 are valid replacing $L(\Sigma(K_{n,n};k))$ by $R(\Sigma(K_{n,n};k))$. This explains why from now on we concentrate exclusively on the derivation of lower bounds such as those of Theorem 5 but concerning $R(\cdot)$.

Our approach partly builds on [17], so to help the reader follow the rest of this article, it will be convenient to have a high level understanding of the proofs of the lower bounds in Theorem 5. We next provide such a general overview. For precise statements and detailed proofs, see [17].

The proof of the lower bound in (3) has two parts, both of which consider a graph G chosen according to $\Sigma(K_{n,n};k)$ whose sides, say A and B, are partitioned into segments $A_1,A_2,\ldots$ and $B_1,B_2,\ldots$, respectively, of roughly the same appropriately chosen size $\widetilde{n}=\widetilde{n}(k)$. For each i, one considers the subgraph of G induced by $A_i\cup B_i$, say G_i, and observes that the union of noncrossing matchings, one for each G_i, is a noncrossing matching of G. The first part of the proof argument is a lower bound on the expected length of a largest noncrossing matching of G_i. The other part of the proof is a lower bound on the expected length of a largest noncrossing matching of G which follows simply by summing the lower bounds from the first part and observing that, by "sub-additivity", $\sum_i L(G_i)\le L(G)$.

Since the size of the segments $A_1,A_2,\ldots$ and $B_1,B_2,\ldots$ is $\widetilde{n}$, there are $n/\widetilde{n}$ such segments in A and in B. An edge of $K_{n,n}$ is in G with probability $1/k$. So the expected number of edges in G_i is $\widetilde{n}^2/k$. The value of $\widetilde{n}$ is chosen so that, for each i, the expected number of edges of G_i is large, and the expected degree of each vertex of G_i is much smaller than 1. Let G'_i be the graph obtained from G_i by removing isolated vertices and the edges incident to vertices of degree greater than 1. By the choice of $\widetilde{n}$, almost all nonisolated vertices of G_i have degree 1. So G'_i has "almost" the same expected number of edges as G_i, i.e., $\widetilde{n}^2/k$ edges. Also, note that G'_i is just a perfect matching (every node has degree exactly 1). This perfect matching, of size say t, defines a permutation of $[t]$ — in fact, by symmetry arguments it is easy to see that, conditioning on t, the permutation is uniformly distributed among all permutations of $[t]$. Observe that a noncrossing matching of G'_i corresponds to an increasing sequence in the aforementioned permutation, and vice versa. So a largest noncrossing matching of G'_i is given by a Longest Increasing Sequence (LIS) of the permutation. There are precise results (by Baik et al. [3]) on the distribution of the length of a LIS of a randomly chosen permutation of $[t]$. The expected length of a LIS for such a random permutation is $2\sqrt{t}$. So a largest noncrossing matching in G'_i has expected length almost $2\sqrt{\widetilde{n}^2/k}=2\widetilde{n}/\sqrt{k}$. As the number of i's is $n/\widetilde{n}$, we obtain a lower bound of almost $(n/\widetilde{n})2\widetilde{n}/\sqrt{k}=2n/\sqrt{k}$ for the expected length of a largest noncrossing matching of G. The same reasoning (although technically significantly more involved) yields a lower tail bound for the deviation of $\sum_i L(G'_i)\le\sum_i L(G_i)\le L(G)$ from $2n/\sqrt{k}$. This concludes our overview of the proof arguments of [17] for deriving the lower bounds of Theorem 5.

We now stress one important aspect of the preceding paragraph discussion. Namely, that by construction G'_i is a subgraph of G_i whose vertices all have degree one, and, moreover, G'_i is in fact an induced subgraph of G. Since G is generated according to $\Sigma(K_{n,n};k)$, it must necessarily be the case that the symbols associated to the edges of G'_i are *all* distinct. Hence, a noncrossing matching of G'_i is also a repetition-free noncrossing matching of G'_i, thence also of G_i. In other words, it holds that $L(G'_i) = R(G'_i) \le R(G_i)$. Thus, a lower tail bound for the deviation of $L(G'_i)$ from $2\widetilde{n}/\sqrt{k}$ is also a lower tail bound for the deviation of $R(G_i)$ from $2\widetilde{n}/\sqrt{k}$. Unfortunately, $R(\cdot)$ is not sub-additive as $L(\cdot)$ above (so now $\sum_i R(G_i)$ is not necessarily a lower bound for $R(G)$). Indeed, the union of repetition-free noncrossing matchings M_i of the G_i's is certainly a noncrossing matching, but is not necessarily repetition-free. This happens because although the symbols associated to the edges of each M_i must be distinct, it might happen that the same symbol is associated to several edges of different M_i's. However, if we can estimate (bound) the number of "symbol overlaps" between edges of distinct M_i's, then we can potentially translate lower tail bounds for the deviation of $L(G'_i) = R(G'_i)$ from some given value, to lower tail bounds for the deviation of $R(G)$ from a properly chosen factor of the given value. This is the approach we will develop in detail in the following sections. However, we still need a lower tail bound for $R(G_i)$ when $\widetilde{n} = \widetilde{n}(k)$ is appropriately chosen in terms of k and G_i is randomly chosen as above. From the previous discussion, it should be clear that such a tail bound is implicitly established in [17]. The formal result of [17] related to G_i, addresses the distribution of $L(\Sigma(K_{r,s};k))$ for $r = s = \widetilde{n}$, as expressed in their Proposition 6 in [17, p. 486]. One can verify that the same result holds, with the same proof, observing that each G_i is distributed according to $\Sigma(K_{\widetilde{n},\widetilde{n}};k)$ and replacing $L(\cdot)$ by $R(\cdot)$. For the sake of future reference, we re-state the claimed result but with respect to the parameter $R(\cdot)$ and the case $r = s = \widetilde{n}$ we are interested in.

Theorem 6. *For every $\delta > 0$, there exists $C = C(\delta)$ such that, if $\widetilde{n}$ is an integer and $C\sqrt{k} \le \widetilde{n} \le \delta k/12$ then, with $m_u = 2(1+\delta)\widetilde{n}/\sqrt{k}$ and $m_l = 2(1-\delta)\widetilde{n}/\sqrt{k}$, for all $t \ge 0$,*

$$\mathbb{P}\big(R(\Sigma(K_{\widetilde{n},\widetilde{n}};k)) \le m_l - t\big) \le 2\, e^{-t^2/8m_u}.$$

4. Distribution of symbols in repetition-free LCSs

One expects that any symbol is equally likely to show up in a repetition-free LCS of two randomly chosen sequences. Intuitively, this follows from the fact that there is a symmetry between symbols. In fact, one expects something stronger to hold; conditioning on the largest repetition-free LCS being of size ℓ, any subset of ℓ symbols among the k symbols of the alphabet should be equally likely. Making this intuition precise is somewhat tricky due to the fact that there might be more than one repetition-free LCS for a given pair of sequences. The purpose of this section is to formalize the preceding discussion.

First, note that if G is in the support of $\Sigma(K_{n,n};k)$, then each of its connected components is either an isolated node or a complete bipartite graph. Hence, each connected component of G is in one-to-one correspondence with a symbol from the k-ary alphabet.

Now, consider some total ordering, denoted $\preceq$, on the noncrossing matchings of $K_{n,n}$. For $\ell \in [k]$, let $\mathcal{G}_\ell$ be the collection of all graphs G in the support of $\Sigma(K_{n,n};k)$ such that $R(G) = \ell$. Given G in $\mathcal{G}_\ell$ let $\mathcal{C}_\ell(G) \subseteq [k]$ denote the collection of symbols assigned to the nodes of the smallest (with respect to the ordering $\preceq$) repetition-free noncrossing matching M of G of size ℓ. Clearly, the cardinality of $\mathcal{C}_\ell(G)$ is ℓ. For G in the support of $\Sigma(K_{n,n};k)$ we say that M is the *canonical* matching of G if M is the smallest, with respect to the ordering $\preceq$, among all largest repetition-free noncrossing matching of G. We claim that, for G chosen according to $\Sigma(K_{n,n};k)$, conditioned on $R(G) = \ell$, the set of symbols associated to the edges of the canonical matching M of G is uniformly distributed over all size ℓ subsets of $[k]$. Formally, we establish the following result.

Lemma 7. *For all $\ell \in [k]$ and $S \subseteq [k]$ with $|S| = \ell$,*

$$\mathbb{P}\big(\mathcal{C}_\ell(G) = S \,|\, R(G) = \ell\big) = \frac{1}{\binom{k}{\ell}},$$

where the probability is taken over the choices of G distributed according to $\Sigma(K_{n,n};k)$.

Proof. For a subset E of edges of $K_{n,n}$, define $\mathcal{P}_\ell(E)$ as the set of elements of $\mathcal{G}_\ell$ whose edge set is exactly E. Let $\mathcal{E}_\ell$ be the collection of all E's such that $\mathcal{P}_\ell(E)$ is nonempty and let $\mathcal{P}_\ell$ be the collection of $\mathcal{P}_\ell(E)$'s where E ranges over subsets of $\mathcal{E}_\ell$. Observe that $\mathcal{P}_\ell$ is a partition of $\mathcal{G}_\ell$. Hence,

$$\sum_{E\in\mathcal{E}_\ell} \mathbb{P}\big(E(G) = E \,|\, R(G) = \ell\big) = \sum_{E\in\mathcal{E}_\ell} \mathbb{P}\big(G \in \mathcal{P}_\ell(E) \,|\, R(G) = \ell\big) = \mathbb{P}\big(G \in \mathcal{G}_\ell \,|\, R(G) = \ell\big) = 1.$$

Moreover,

$$\mathbb{P}\big(\mathcal{C}_\ell(G) = S \,|\, R(G) = \ell\big) = \sum_{E\in\mathcal{E}_\ell} \mathbb{P}\big(\mathcal{C}_\ell(G) = S \,|\, E(G) = E\big)\mathbb{P}\big(E(G) = E \,|\, R(G) = \ell\big).$$

Thus, the desired conclusion will follow once we show that $\mathbb{P}\big(\mathcal{C}_\ell(G) = S \,|\, E(G) = E\big) = 1/\binom{k}{\ell}$ for all $E \in \mathcal{E}_\ell$. Indeed, let $E \in \mathcal{E}_\ell$ and observe that the condition $E(G) = E$ uniquely determines the canonical noncrossing matching of G of size ℓ, say $M = M(G)$. Moreover, note that any choice of distinct ℓ symbols to each of the ℓ distinct components of G to which the edges of M belong is equally likely. Since there are $\binom{k}{\ell}$ possible choices of ℓ-symbol subsets of $[k]$, the desired conclusion follows. □

The preceding result will be useful in the next section in order to address the following issue. For G and $G_1, \ldots, G_b$ as defined in Section 3, suppose that $M_1, \ldots, M_b$ are the largest repetition-free noncrossing matchings of $G_1, \ldots, G_b$, respectively. As mentioned before, the union M of the M_i's is a noncrossing matching of G, but not necessarily repetition-free. Obviously, we can remove edges from M, keeping one edge for each symbol associated to the edges of M, and thus obtain a repetition-free noncrossing matching M' contained in M, and thence also in G. Clearly, it is of interest to determine the expected number of edges that are removed from M to obtain M', i.e., $|M \setminus M'|$, and in particular whether this number is small. Lemma 7 is motivated, and will be useful, in this context. The reason being that, conditioning on the size s_i of the largest repetition-free noncrossing matching in each G_i, it specifies the distribution of the set of symbols $\mathcal{C}_{s_i}(G_i)$ associated to the edges of the canonical noncrossing matching of G_i. The latter helps in the determination of the sought-after expected value, since

$$|M \setminus M'| = \sum_{i=1}^{b} \left|\mathcal{C}_{s_i}(G_i)\right| - \left|\cup_{i=1}^{b} \mathcal{C}_{s_i}(G_i)\right|.$$

5. Tail bounds

In this section we derive bounds on the probability that $R(G)$ is bounded away from its expected value when G is chosen according to $\Sigma(K_{n,n}; k)$. We will ignore the case where $n = O(\sqrt{k})$ due to its limited interest and the impossibility of deriving meaningful asymptotic results. Indeed, if $n \le C\sqrt{k}$ for some positive constant C and sufficiently large k, then the expected number of edges of a graph G chosen according to $\Sigma(K_{n,n}; k)$ is $n^2/k \le C^2$ (just observe that there are n^2 potential edges and that each one occurs in G with probability $1/k$). Since $0 \le R(G) \le |E(G)|$, when $n = O(\sqrt{k})$, the expected length of a repetition-free LCS will be constant — hence, not well suited for an asymptotic study. Thus, we henceforth assume that $n = \omega(\sqrt{k})$. If in addition $n = o(k)$, then Theorem 6 already provides the type of tail bounds we are looking for. Hence, we need only consider the case where $n = \Omega(k)$. We will show that three different regimes arise. The first one corresponds to $n = o(k\sqrt{k})$. For this case we show that the length of a repetition-free LCS is concentrated around its expected value, which in fact is roughly $2n/\sqrt{k}$ (i.e., the same magnitude as that of the length of a standard LCS). The second one corresponds to $n = \Theta(k\sqrt{k})$. For this regime we show that the length of a repetition-free LCS cannot be much smaller than a fraction of k, and we relate the constant of proportionality with the constant hidden in the asymptotic dependency $n = \Theta(k\sqrt{k})$. The last regime corresponds to $n = (1 + \Omega(1))k\sqrt{k} \ln k$. For this latter case we show that with high probability a repetition-free LCS is of size k.

Throughout this section, n and k are positive integers, G is a bipartite graph chosen according to $\Sigma(K_{n,n}; k)$, and $G_1, \ldots, G_b$ are as defined in Section 3, where b is an integer approximately equal to $n/\widetilde{n}$. Note in particular that G_i is distributed according to $\Sigma(K_{\widetilde{n},\widetilde{n}}; k)$.

This section's first formal claim is motivated by an obvious fact; if $r = R(G)$ is "relatively small", then at least one of the two following situations must happen:

- For some $i \in [b]$, the value of $r_i = R(G_i)$ is "relatively small".
- The sets of symbols, $\mathcal{C}_{r_i}(G_i)$, associated to the edges of the canonical largest noncrossing matching of G_i, for $i \in [b]$, have a "relatively large" overlap, more precisely, the cardinality of $\mathcal{C}_r(G)$ is "relatively small" compared to the sum, for $i \in [b]$, of the cardinalities of $\mathcal{C}_{r_i}(G_i)$.

The next result formalizes the preceding observation. In particular, it establishes that the probability that $R(G)$ is "relatively small" is bounded by the probability that one of the two aforementioned cases occurs (and also gives a precise interpretation to the terms "relatively large/small").

Lemma 8. *Let b be a positive integer. For $a \ge 0$ and $r \ge t \ge 0$, let*

$$P_1 = P_1(r, t) \stackrel{\text{def}}{=} \sum_{\substack{r_1,\ldots,r_b \ge 0 \\ r_1+\cdots+r_b=\lfloor r-t \rfloor}} \mathbb{P}\big(R(G_i) \le r_i, \forall i \in [b]\big), \quad \textit{(Definition of } P_1\textit{)}$$

$$P_2 = P_2(a, r, t) \stackrel{\text{def}}{=} \mathbb{P}\left(R(G) \le r - a, \sum_{i=1}^{b} R(G_i) \ge r - t\right). \quad \textit{(Definition of } P_2.\textit{)}$$

Then,

$$\mathbb{P}\big(R(G) \le r - a\big) \le P_1 + P_2.$$

Proof. Just note that

$$\mathbb{P}\big(R(G) \le r-a\big) = \mathbb{P}\Big(R(G) \le r-a, \sum_{i=1}^{b} R(G_i) < \lceil r-t \rceil\Big) + \mathbb{P}\Big(R(G) \le r-a, \sum_{i=1}^{b} R(G_i) \ge \lceil r-t \rceil\Big)$$
$$\le \mathbb{P}\Big(\sum_{i=1}^{b} R(G_i) < r-t\Big) + \mathbb{P}\Big(R(G) \le r-a, \sum_{i=1}^{b} R(G_i) \ge r-t\Big).$$

The desired conclusion follows observing that the two last terms in the preceding displayed expression are equal to P_1 and P_2, respectively. □

The following lemma will be useful in bounding the terms in P_1, i.e., the probability that $R(G_i)$ is "relatively small" for some i. Henceforth, for the sake of clarity of exposition, we will ignore the issue of integrality of quantities (since we are interested in the case where n is large, ignoring integrality issues should have a negligible and vanishing impact in the following calculations).

Lemma 9. *Let $\delta > 0$ and $\widetilde{n} = \widetilde{n}(k)$ be such that it satisfies the hypothesis of Theorem 6 and let $m_l = (1-\delta)2\widetilde{n}/\sqrt{k}$. Let $b = b(k) \stackrel{\text{def}}{=} n/\widetilde{n}$. Then, for every $t \ge 0$*

$$P_1 = P_1(bm_l, t) \le (2e(m_l+1))^b \exp\Big(-\frac{t^2}{16(1+\delta)n/\sqrt{k}}\Big).$$

Proof. Observe that, by independence of the $R(G_i)$'s and Theorem 6, for $m_u = (1+\delta)2\widetilde{n}/\sqrt{k}$,

$$\mathbb{P}\big(R(G_i) \le r_i, \forall i \in [b]\big) = \mathbb{P}\big(R(G_i) \le m_l - (m_l - r_i), \forall i \in [b]\big)$$
$$\le \prod_{i=1}^{b} (2e^{-\max\{0, m_l - r_i\}^2/8m_u}) = 2^b e^{-\sum_{i=1}^{b} \max\{0, m_l - r_i\}^2/8m_u}.$$

By Cauchy–Schwarz, since $\max\{0,x\} + \max\{0,y\} \ge \max\{0, x+y\}$, and assuming that for some given $t \ge 0$ it holds that $\sum_{i=1}^{b} r_i = \lfloor bm_l - t \rfloor$,

$$\sum_{i=1}^{b} \max\{0, m_l - r_i\}^2 \ge \frac{1}{b}\Big(\sum_{i=1}^{b} \max\{0, m_l - r_i\}\Big)^2 \ge \frac{1}{b} t^2.$$

Recalling that there are $\binom{M+b-1}{b-1} \le \binom{M+b}{b}$ ways in which b nonnegative summands can add up to $M \in \mathbb{N}$, and that $\binom{\mu}{\nu} \le (e\mu/\nu)^{\nu}$,

$$P_1 \le \sum_{\substack{r_1,\ldots,r_b \ge 0 \\ r_1+\cdots+r_b = \lfloor bm_l - t \rfloor}} 2^b e^{-t^2/8bm_u} \le \binom{\lfloor bm_l - t \rfloor + b}{b} 2^b e^{-t^2/8bm_u} \le (2e(m_l+1))^b e^{-t^2/8bm_u}.$$

Since $m_u = (1+\delta)2\widetilde{n}/\sqrt{k}$ and $b\widetilde{n} = n$, the desired conclusion follows immediately. □

The next lemma will be useful in bounding P_2, i.e., the probability that the sets of symbols associated to the edges of the canonical largest noncrossing G_i's matchings have a "relatively large" overlap. The result in fact shows how to translate tail bounds for an urn occupancy model into bounds for P_2.

Lemma 10. *If b is a positive integer, $a \ge 0$, $r \ge t \ge 0$, and $s = \lceil r - t \rceil$, then*

$$P_2 = P_2(a, r, t) \le \mathbb{P}\big(k - Y^{(k,s)} \le r - a\big).$$

Proof. Clearly,

$$P_2 = \sum_{\substack{s_1,\ldots,s_b \ge 0 \\ s_1+\cdots+s_b \ge r-t}} \mathbb{P}\big(R(G) \le r-a \,|\, R(G_i) = s_i, \forall i \in [b]\big)\, \mathbb{P}\big(R(G_i) = s_i, \forall i \in [b]\big).$$

Let $\mathcal{C}_\ell(\cdot)$ be as defined in Section 4. Note that if we take the union of noncrossing matchings, one M_i for each G_i, we get a noncrossing matching $M = \cup_i M_i$ of G. However, the edges of M do not necessarily have distinct associated symbols. By throwing away all but one of the edges of M associated to a given symbol, one obtains a repetition-free noncrossing matching of G. It follows that, conditioning on $R(G_i) = s_i$ for all $i \in [b]$,

$$R(G) \ge \Big|\bigcup_{i=1}^{b} \mathcal{C}_{s_i}(G_i)\Big|.$$

Thus,

$$\begin{aligned}\mathbb{P}\big(R(G) \le r-a \,|\, R(G_i)=s_i, \forall i \in [b]\big) &\le \mathbb{P}\bigg(\Big|\bigcup_{i=1}^{b} \mathcal{C}_{s_i}(G_i)\Big| \le r-a \,|\, R(G_i)=s_i, \forall i \in [b]\bigg)\\ &= \mathbb{P}\bigg(\Big|\bigcup_{i=1}^{b} \mathcal{C}_{s_i}(G_i)\Big| \le r-a \,\Big|\, \big|\mathcal{C}_{s_i}(G_i)\big| = s_i, \forall i \in [b]\bigg).\end{aligned}$$

Let $\vec{s} = (s_1, \dots, s_b)$. We claim that $\big|\bigcup_{i=1}^{b} \mathcal{C}_{s_i}(G_i)\big|$ conditioned on $\big|\mathcal{C}_{s_i}(G_i)\big| = s_i$, for all $i \in [b]$, is distributed exactly as the number of nonempty urns left when the Grouped Urn $(k, \vec{s})$-model (as defined in Section 2) ends, i.e., is distributed as the random variable $k - X^{(k,\vec{s})}$ (where $X^{(k,\vec{s})}$ is as defined in Section 2). Indeed, it suffices to note that by Proposition 4, conditioned on $\big|\mathcal{C}_{s_i}(G_i)\big| = s_i$, the set $S_i = \mathcal{C}_{s_i}(G_i)$ is a randomly and uniformly chosen subset of $[k]$ of size s_i, and that $k - X^{(k,\vec{s})}$ is distributed exactly as $\big|\bigcup_{i=1}^{b} \mathcal{C}_{s_i}(G_i)\big|$. It follows, from the forgoing discussion and Lemma 3, that

$$\mathbb{P}\big(R(G) \le r-a \,|\, R(G_i)=s_i, \forall i \in [b]\big) = \mathbb{P}\big(k - X^{(k,\vec{s})} \le r-a\big) \le \mathbb{P}\big(k - Y^{(k,s)} \le r-a\big). \quad \square$$

The next result establishes the first of the announced tail bounds, for the first of the three regimes indicated at the start of this section. An interesting aspect, that is not evident from the theorem's statement, is the following fact that is implicit in its proof; if the speed of growth of n as a function of k is not too fast, then we may choose b as a function of k so that $\sum_{i=1}^{b} R(G_i)$ is roughly (with high probability) equal to $R(G)$. In particular, the proof argument rests on the fact that, for an appropriate choice of parameters, the canonical largest noncrossing matching of G_i is of size approximately $2(n/b)/\sqrt{k}$, and with high probability there is very little overlap between the symbols associated to the edges of the canonical largest noncrossing matchings of distinct G_i's.

Proposition 11. *If $n = o(k\sqrt{k})$, then for every $0 < \xi \le 1$ there is a constant $k_0 = k_0(\xi)$ such that, for all $k > k_0$,*

$$\mathbb{P}\big(R(G) \le (1-\xi)2n/\sqrt{k}\big) \le 2e^{-\frac{1}{10}\xi^2 2n/\sqrt{k}}.$$

Proof. Let $c > 1$ be large enough so $(1-1/c)^2 \ge (9/10)(1+\xi/c)$. Let $\delta = \xi/c$ and $t = (1-1/c)\xi 2n/\sqrt{k}$. Now, choose $\widetilde{n} = \widetilde{n}(k) = k^{3/4}$ (instead of 3/4, any exponent strictly between 1/2 and 1 suffices). Note that one can choose $\widetilde{k}_0$ (depending on ξ through δ) so that for all $k \ge \widetilde{k}_0$ the conditions on $\widetilde{n}$ of Theorem 6 are satisfied. Let m_l and m_u be as in Theorem 6. Note that $m_l = (1-\xi/c)2\tilde{n}/\sqrt{k} = \Theta(k^{1/4})$, $b = n/\tilde{n} = n/k^{3/4}$, and $bm_u = (1+\xi/c)2n/\sqrt{k}$. Hence, by Lemma 9,

$$P_1 \le \exp\Big(b\ln(2e(m_l+1)) - \frac{t^2}{8bm_u}\Big) = \exp\Big(\frac{n}{\sqrt{k}}\Theta\big(k^{-1/4}\ln k\big) - \frac{(1-1/c)^2\xi^2 2n/\sqrt{k}}{8(1+\xi/c)}\Big).$$

Since $k^{-1/4}\ln k = o(1)$ and $(1-1/c)^2 \ge (9/10)(1+\xi/c)$, it follows that for a sufficiently large $k_0' \ge \widetilde{k}_0$ it holds that for all $k \ge k_0'$,

$$P_1 \le \exp\Big(-\frac{(1-1/c)^2\xi^2 2n/\sqrt{k}}{9(1+\xi/c)}\Big) \le \exp\Big(-\frac{1}{10}\xi^2 2n/\sqrt{k}\Big).$$

On the other hand, since $t = (1-1/c)\xi 2n/\sqrt{k}$, if we fix $a = \xi 2n/\sqrt{k}$, then we have that $t - a = -(\xi/c)2n/\sqrt{k}$. Taking $s = bm_l - t = (1-\xi)2n/\sqrt{k} \le 2n/\sqrt{k}$, as $\xi \le 1$, by Lemma 10 and Proposition 4, Part 3,

$$P_2 \le \mathbb{P}\big(k - Y^{(k,s)} \le bm_l - a\big) = \mathbb{P}\big(k - Y^{(k,s)} \le s + t - a\big) \le \mathbb{P}\bigg(k - Y^{(k,s)} \le s - \frac{\xi 2n}{c\sqrt{k}}\bigg) \le \Big(\frac{2cen}{\xi k\sqrt{k}}\Big)^{(\xi/c)2n/\sqrt{k}}.$$

Let k_0'' be sufficiently large (depending on ξ) so that $2cen/(\xi k\sqrt{k}) \le e^{-c\xi/10}$ for all $k \ge k_0''$ (such a k_0'' exists because $n = o(k\sqrt{k})$). It follows that for $k \ge k_0''$ we can upper bound P_2 by $\exp\big(-\frac{1}{10}\xi^2 2n/\sqrt{k}\big)$.

Since by Lemma 8 we know that $\mathbb{P}\big(R(G) \le (1-\xi)2n/\sqrt{k}\big) \le P_1 + P_2$, it follows that for $k \ge k_0 = k_0(\xi) \stackrel{\text{def}}{=} \max\{k_0', k_0''\}$ we get the claimed bound. $\square$

Next, we consider a second regime, but first we establish an inequality that we will soon apply.

Claim 12. *For every $0 \le x \le 1$ and $\rho \ge 0$,*

$$e^{-\rho(1-x)} - e^{-\rho} - x(1 - e^{-\rho}) \le 0.$$

Proof. Since $0 \le x \le 1$, it holds that $0 \le x^n \le x$ for all $n \in \mathbb{N} \setminus \{0\}$. Then, since $e^y = \sum_{n\in\mathbb{N}} y^n/n!$ and performing some basic arithmetic,

$$e^{\rho x} - 1 - x(e^{\rho} - 1) = \sum_{n\ge 1} \frac{\rho^n}{n!}(x^n - x) \le 0.$$

Multiplying by $e^{-\rho}$, the claimed result immediately follows. □

Proposition 13. *Let $\rho > 0$ and $0 < \xi < 1$. If $n = \frac{1}{2}\rho k\sqrt{k}$, then there is a sufficiently large constant $k_0 = k_0(\rho, \xi)$ such that, for all $k > k_0$,*

$$\mathbb{P}\big(R(G) \le (1-\xi)k(1-e^{-\rho})\big) \le 2e^{-\frac{\xi^2}{32(1+\xi/12)}k(1-e^{-\rho})} \le 2e^{-\frac{1}{35}\xi^2 k(1-e^{-\rho})}.$$

Proof. Since $\xi < 1$, the second stated inequality follows immediately from the first one. We thus focus on establishing the first stated inequality.

Let $\delta = \xi/12$. Now, choose $\tilde{n} = k^{3/4}$ (instead of 3/4, any exponent strictly between 1/2 and 1 suffices) and set $b = n/\tilde{n}$. Note that one can choose k_0' (depending on ξ through δ) so that for all $k > k_0'$ the conditions on $\tilde{n} = \tilde{n}(k)$ of Theorem 6 are satisfied. Let $m_l = (1-\delta)2\tilde{n}/\sqrt{k}$ and observe that $bm_l = (1-\delta)2n/\sqrt{k} = (1-\xi/12)\rho k$. Choose $t = (2\xi/3)\rho k$ and note that $s = bm_l - t = (1-3\xi/4)\rho k$. Let $\lambda = \mathbb{E}\big(Y^{(k,s)}\big) = k(1-1/k)^s$ be as in Proposition 4. We claim that, for $0 < \xi < 1$,

$$\lambda \le ke^{-\rho} + (3\xi/4)k(1-e^{-\rho}). \tag{4}$$

Indeed, since $1 + x \le e^x$ we have that $\lambda = k(1-1/k)^s \le ke^{-\rho(1-3\xi/4)}$, so to prove (4) it suffices to recall that by Claim 12 we have that $e^{-\rho(1-3\xi/4)} \le e^{-\rho} + (3\xi/4)(1-e^{-\rho})$.

Now, fix $\tilde{a}$ and a so $\tilde{a} = \xi k(1-e^{-\rho})$ and $a = bm_l - k(1-e^{-\rho}) + \tilde{a}$. By Lemma 10 and (4),

$$P_2 \le \mathbb{P}\big(k - Y^{(k,s)} \le bm_l - a\big) = \mathbb{P}\big(Y^{(k,s)} \ge ke^{-\rho} + \tilde{a}\big) \le \mathbb{P}\big(Y^{(k,s)} \ge \lambda + (\xi/4)k(1-e^{-\rho})\big).$$

Hence, taking $p = \lambda/k \le 1$, $q = 1-p$, and applying Proposition 4, Part 1,

$$P_2 \le \exp\Big(-\frac{\xi^2(1-e^{-\rho})^2 k}{32(pq + (\xi/12)(1-e^{-\rho}))}\Big) \le \exp\Big(-\frac{\xi^2(1-e^{-\rho})^2 k}{32(q + (\xi/12)(1-e^{-\rho}))}\Big).$$

Again by Proposition 4, we know that $\lambda = k(1-1/k)^s$. Thus, recalling that by our choice of parameters $s = (1-3\xi/4)\rho k$ and since $(1-1/k)^s = (1-1/k)^{\rho(1-3\xi/4)k}$ converges to $e^{-\rho(1-3\xi/4)} > e^{-\rho}$ when k goes to ∞, it follows that $q = 1-p = 1-(1-1/k)^s$ can be upper bounded by $1-e^{-\rho}$ for all $k > k_0''$ and some sufficiently large $k_0'' > k_0'$ (depending on ξ). Hence, for $k > k_0''$ it holds that $q + (\xi/12)(1-e^{-\rho}) \le (1+\xi/12)(1-e^{-\rho})$, and

$$P_2 \le \exp\Big(-\frac{\xi^2}{32(1+\xi/12)}k(1-e^{-\rho})\Big).$$

We will now upper bound P_1. Note that, by our choice for $\tilde{n}$ and the hypothesis on n, we have $m_l = (1-\xi/12)2\tilde{n}/\sqrt{k} = (1-\xi/12)k^{1/4} \le k$, $b = n/\tilde{n} \le \rho k^{3/4}$, and $bm_u = (1+\xi/12)2n/\sqrt{k} = (1+\xi/12)\rho k$. Recalling that we fixed $t = (2\xi/3)\rho k$, by Lemma 9,

$$P_1 \le \exp\Big(b\ln(2e(m_l+1)) - \frac{t^2}{8bm_u}\Big) \le \exp\Big(\rho k^{3/4}\ln(2e(k+1)) - \frac{\xi^2\rho k}{18(1+\xi/12)}\Big).$$

Since $k^{3/4}\ln(2e(k+1)) = o(k)$ and because $1-e^{-\rho} \le \rho$, it follows that for a sufficiently large k_0''' it holds that, for all $k > k_0'''$,

$$P_1 \le \exp\Big(-\frac{\xi^2}{19(1+\xi/12)}\rho k\Big) \le \exp\Big(-\frac{\xi^2}{19(1+\xi/12)}k(1-e^{-\rho})\Big).$$

Since $\mathbb{P}\big(R(G) \le (1-\xi)k(1-e^{-\rho})\big) \le P_1 + P_2$ for $k > k_0 = k_0(\rho,\xi) \stackrel{\text{def}}{=} \max\{k_0'', k_0'''\}$, we get the claimed bound. □

Our next result establishes that if n is sufficiently large with respect to k, then with high probability a repetition-free LCS is of size k, i.e., it is a permutation of the underlying alphabet. Moreover, the theorem's proof implicitly shows something stronger; if the speed of growth of n as a function of k is fast enough, then we may choose b as a function of k so that with high probability every symbol of the k-ary alphabet shows up in association to an edge of a canonical maximum size matching of some G_i — chosen such edges, one obtains a noncrossing repetition free matching of G of the maximum possible size k.

Proposition 14. *If $n = (\frac{1}{2}+\xi)k\sqrt{k}\ln k$ for some $\xi > 0$, then there is a constant $k_0 = k_0(\xi)$ such that, for all $k > k_0$,*

$$\mathbb{P}\big(R(G) \ne k\big) \le \frac{2}{k^{\xi}}.$$

Proof. Let $\delta = \delta(\xi) > 0$ be such that $(1-\delta)(1+2\xi) = 1+3\xi/2$. Now, let $\tilde{n} = k^{3/4}$ (instead of 3/4, any exponent strictly between 1/2 and 1 suffices) and set $b = n/\tilde{n}$. Note that one can choose k_0' (depending on ξ through δ) so that for all $k > k_0'$ the conditions on $\widetilde{n} = \widetilde{n}(k)$ of Theorem 6 are satisfied. Let $m_l = (1-\delta)2\widetilde{n}/\sqrt{k}$ be as in Theorem 6. Observe that $bm_l = (1-\delta)2n/\sqrt{k} = (1+3\xi/2)k\ln k$. Choose $t = (\xi/2)k\ln k$ so that $s = bm_l - t = (1+\xi)k\ln k$. Fix a so $k - bm_l + a = 1$. By Lemma 10 and Proposition 4, Part 2,

$$P_2 \le \mathbb{P}\big(Y^{(k,s)} \ge k - bm_l + a\big) = \mathbb{P}\big(Y^{(k,s)} \ne 0\big) \le \frac{1}{k^{\xi}}.$$

By the hypothesis on n and the choice of $\widetilde{n}$, we have that $b = n/\tilde{n} = (\frac{1}{2}+\xi)k^{3/4}\ln k$, so recalling that $m_l = (1-\delta)2\widetilde{n}/\sqrt{k} = (1-\delta)2k^{1/4}$,

$$b\ln(2e(m_l+1)) = O(k^{3/4}\ln^2 k) = o(k\ln k).$$

Furthermore, let $m_u = (1+\delta)2\widetilde{n}/\sqrt{k}$ be as in Theorem 6. Thus,

$$\frac{t^2}{16(1+\delta)n/\sqrt{k}} = \frac{t^2}{8bm_u} = \frac{\xi^2 k\ln k}{32(1+\delta)(1+2\xi)}.$$

Hence, by Lemma 9, for a sufficiently large constant k_0'' (again depending on ξ through δ), we can guarantee that, for all $k > k_0''$,

$$P_1 \le (2e(m_l+1))^b e^{-t^2/8bm_u} = \exp\Big(b\ln(2e(m_l+1)) - \frac{t^2}{8bm_u}\Big) \le \frac{1}{k^{\xi}}.$$

Summarizing, for $k > k_0 = k_0(\xi) \stackrel{\text{def}}{=} \max\{k_0', k_0''\}$, we get that $\mathbb{P}\big(R(G) \ne k\big) \le P_1 + P_2 \le 2/k^{\xi}$. $\square$

From the lower tail bounds for $R(G)$ obtained above, one can easily derive lower bounds on the expected value of $R(G)$ via the following well-known trick.

Lemma 15. *If X is a nonnegative random variable and $x > 0$, then $\mathbb{E}\big(X\big) \ge x(1 - \mathbb{P}\big(X \le x\big))$.*

Proof. Let $\mathbb{I}_A$ denote the indicator of the event A occurring. Just observe that

$$\mathbb{E}\big(X\big) = \mathbb{E}\big(X\mathbb{I}_{\{X\le x\}}\big) + \mathbb{E}\big(X\mathbb{I}_{\{X>x\}}\big) \ge x\,\mathbb{E}\big(\mathbb{I}_{\{X>x\}}\big) = x\big(1 - \mathbb{P}\big(X \le x\big)\big). \quad \square$$

Theorem 2 encompasses Propositions 11, 13 and 14. Theorem 1 follows as a direct consequence of Theorem 2, the preceding lemma, and the fact that $R(G) \le k$.

6. Final comments

Before concluding, we discuss a byproduct of our analysis. To do so, we note that the computational experiments presented by Adi et al. [1] considered problem instances where sequences of n symbols where randomly, uniformly, and independently chosen over a k-ary alphabet. The experimental findings are consistent with our estimates of $\mathbb{E}\big(R_n\big)$ (recall R_n was defined in the introduction). Our results thus have the added bonus, at least when n and k are large, that they allow to perform comparative studies, as the aforementioned one, but replacing the (expensive) exact computation of R_n by our estimated value. Our analysis also suggests that additional experimental evaluation of proposed heuristics, over test cases generated as in so called *planted random models*, might help to further validate the usefulness of proposed algorithmic approaches. Specifically, for the RFLCS problem, according to the planted random model, one way to generate test cases would be as described next. First, for some fixed $\ell \le k$, choose a repetition-free sequence z of length $\ell < n$ over a k-ary alphabet. Next, generate a sequence x' of n symbols randomly, uniformly, and independently over the k-ary alphabet. Finally, uniformly at random choose a size ℓ collection $s_1, \ldots, s_\ell \subseteq \{1, \ldots, n\}$ of distinct positions of x' and replace the s_ith symbol of x' by the ith symbol of z, thus "planting" z in x'. Let x be the length n sequence thus obtained. Repeat the same procedure again for a second sequence y' also of n randomly chosen symbols but with the same sequence z, and obtain a new sequence y. The resulting sequences x and y are such that $\mathrm{RFLCS}(x, y) \ge \ell$. The parameter ℓ can be chosen to be larger than the value our study predicts for R_n. This allows to efficiently generate "non typical" problem instances over which to try out the heuristics, as well as a lower bound certificate for the problem optimum (although, not a matching upper bound). For more details on the planted random model the interested reader is referred to Bui, Chaudhuri, Leighton, and Sipser [9], where (to the best of our knowledge) the model first appeared, and to follow up work by Boppana [8], Jerrum and Sorkin [16], Condon and Karp [12], and the more recent paper of Coja-Oghlan [11].

Acknowledgments

The authors would like to thank Carlos E. Ferreira, Yoshiharu Kohayakawa, and Christian Tjandraatmadja for some discussions in the preliminary stages of their investigation.

The first author was partially supported by CNPq 308523/2012-1, 477203/2012-4 and Proj. MaCLinC of NUMEC/USP. The second author gratefully acknowledges the support of Millennium Nucleus Information and Coordination in Networks ICM/FIC RC130003 and CONICYT via Basal in Applied Mathematics.

References

[1] S. Adi, M. Braga, C. Fernandes, C. Ferreira, F. Martinez, M.-F. Sagot, M. Stefanes, C. Tjandraatmadja, Y. Wakabayashi, Repetition-free longest common subsequence, Discrete Appl. Math. 158 (12) (2010) 1315–1324.

[2] S. Angibaud, G. Fertin, I. Rusu, A. Thvenin, S. Vialette, On the approximability of comparing genomes with duplicates, J. Graph Algorithms Appl. 13 (1) (2009) 19–53.

[3] J. Baik, P. Deift, K. Johansson, On the distribution of the length of the longest increasing subsequence of random permutations, J. Amer. Math. Soc. 12 (1999) 1119–1178.

[4] L. Bergroth, H. Hakonen, T. Raita, A survey of longest common subsequence algorithms, in: Proceedings of the 7th International Symposium on String Processing Information Retrieval, SPIRE, 2000, pp. 39–48.

[5] G. Blin, P. Bonizzoni, R. Dondi, F. Sikora, On the parameterized complexity of the repetition free longest common subsequence problem, Inform. Process. Lett. 112 (7) (2012) 272–276.

[6] P. Bonizzoni, G. Della Vedova, R. Dondi, G. Fertin, R. Rizzi, S. Vialette, Exemplar longest common subsequence, IEEE/ACM Trans. Comput. Biol. Bioinform. 4 (4) (2007) 535–543.

[7] P. Bonizzoni, G. Della Vedova, R. Dondi, Y. Pirola, Variants of constrained longest common subsequence, Inform. Process. Lett. 110 (20) (2010) 877–881.

[8] R. Boppana, Eigenvalues and graph bisection: An average-case analysis, in: Proceedings of the 28th Annual Symposium on Foundations of Computer Science, FOCS, IEEE Computer Society, 1987, pp. 280–285.

[9] T. Bui, S. Chaudhuri, T. Leighton, M. Sipser, Graph bisection algorithms with good average case behavior, Combinatorica 7 (2) (1987) 171–191.

[10] M. Castelli, S. Beretta, L. Vanneschi, A hybrid genetic algorithm for the repetition free longest common subsequence problem, Oper. Res. Lett. 41 (6) (2013) 644–649.

[11] A. Coja-Oghlan, A spectral heuristic for bisecting random graphs, Random Structures Algorithms 29 (3) (2006) 351–398.

[12] A. Condon, R. Karp, Algorithms for graph partitioning on the planted partition model, Random Structures Algorithms 18 (2) (2001) 116–140.

[13] W. Feller, An introduction to Probability Theory and its Applications, Vol. 1, third edition, John Wiley & Sons, 1968.

[14] C.E. Ferreira, C. Tjandraatmadja, A branch-and-cut approach to the repetition-free longest common subsequence problem, Electron. Notes Discrete Math. 36 (2010) 527–534.

[15] S. Janson, Large Deviation Inequalities for Sums of Indicator Variables, Tech. Rep. 1994:34, Uppsala Universitet, 1994, Available at http://www2.math.uu.se/~svante/papers/sj107.ps.

[16] M. Jerrum, G.B. Sorkin, Simulated annealing for graph bisection, in: Proceedings of the 34th Annual Symposium on Foundations of Computer Science, FOCS, IEEE Computer Society, 1993, pp. 94–103.

[17] M. Kiwi, M. Loebl, J. Matoušek, Expected length of the longest common subsequence for large alphabets, Adv. Math. 197 (2005) 480–498.

[18] D. Sankoff, Genome rearrangement with gene families, Bioinformatics 15 (11) (1999) 909–917.

[19] D. Sankoff, Gene and genome duplication, Curr. Opin. Genet. Dev. 11 (6) (2001) 681–684.

[20] D. Sankoff, J. Kruskal (Eds.), Common Subsequences and Monotone Subsequences, Addison–Wesley, Reading, Mass, 1983, pp. 363–365. (Chapter 17).

[21] J.M. Steele, Probability Theory and Combinatorial Optimization, in: CBMS-NSF Regional Conference Series in Applied Mathematics, SIAM, 1996.

[22] S. Vialette, Algorithmic Contributions to Computational Molecular Biology. Data Structures and Algorithms, Université Paris-Est, 2010, Available at https://tel.archives-ouvertes.fr/tel-00862069.