Ann. Comb. 27 (2023) 749–768

Published online May 23, 2023
https://doi.org/10.1007/s00026-023-00651-5

Annals of Combinatorics

Long Twins in Random Words

Andrzej Dudek, Jarosław Grytczuk and Andrzej Ruciński

Abstract. *Twins* in a finite word are formed by a pair of identical subwords placed at disjoint sets of positions. We investigate the maximum length of twins in *a random* word over a k-letter alphabet. The obtained lower bounds for small values of k significantly improve the best estimates known in the deterministic case. Bukh and Zhou in 2016 showed that every ternary word of length n contains twins of length at least $0.34n$. Our main result states that in a random ternary word of length n, with high probability, one can find twins of length at least $0.41n$. In the general case of alphabets of size $k \geqslant 3$ we obtain analogous lower bounds of the form $\frac{1.64}{k+1}n$ which are better than the known deterministic bounds for $k \leqslant 354$. In addition, we present similar results for *multiple* twins in random words.

1. Introduction

Looking for twin objects has had a long tradition in all branches of mathematics. For discrete structures a first question of this kind is attributed to Ulam (see, e.g., [7]) who proposed to measure similarity of two graphs in terms of their edge decompositions into pairwise isomorphic subgraphs. If the number of subgraphs in a decomposition is small, then there must be a large pair among them, which leads to the following general question: given a combinatorial structure (or a pair of structures), how large disjoint isomorphic substructures does it (do they) contain?

This problem has been studied in various forms: for graphs [1,16], words [3,6], and permutations [5,8–10]. Lots of interesting results and challenging open problems can be found in these papers and their references. Some of them are collected in a survey [2].

A. Dudek was supported in part by Simons Foundation Grant #522400. J. Grytczuk was supported in part by Narodowe Centrum Nauki, Grant 2020/37/B/ST1/03298. A. Ruciński was supported in part by Narodowe Centrum Nauki, Grant 2018/29/B/ST1/00426.

1.1. Twins in Words: History and Background

In this paper, we study the problem of twins in random words. To put our work in a broader context, we briefly recall what is known on twins in words in general.

The first who introduced and studied this problem were Axenovich, Person, and Puzynina [3]. We say that a word $w = w_1w_2\cdots w_n \in A^n$, $|A| < \infty$, contains *twins* of *length* t if there exist *disjoint* subsets $\{i_1 < i_2 < \cdots < i_t\}$ and $\{j_1 < j_2 < \cdots < j_t\}$ of indices such that $w_{i_1}w_{i_2}\cdots w_{i_t} = w_{j_1}w_{j_2}\cdots w_{j_t}$. Let $t(w)$ denote the maximum length of twins in w and let $t(n,k)$ denote the minimum of $t(w)$ over all words w of length n over a k-element alphabet A. They proved in [3] that $t(n,2) = n/2 - o(n)$. In other words, every finite binary word can be split into a pair of identical subwords, up to an asymptotically negligible remainder. To get this striking result they developed a regularity lemma for words—an interesting tool with a great potential for further applications (see, e.g., [12]).

At present, it is not known whether the same is true for alphabets of size 3. By considering the sub-word formed by two most frequent letters, the above mentioned estimate $t(n,2) = n/2 - o(n)$ from [3] implies that $t(n,k) \geqslant n/k - o(n)$. For $k \geqslant 3$, this was slightly improved by Bukh and Zhou [6] to

$$t(n,k) \geqslant 1.02 \cdot \frac{n}{k} - o(n) \tag{1.1}$$

and to

$$t(n,k) \geqslant \left(\frac{k}{81}\right)^{1/3} \cdot \frac{n}{k} - (k/3)^{1/3}, \tag{1.2}$$

the latter being a (much) better estimate for larger k. In [3,6] there are also some upper bounds on $t(n,k)$ valid for $k \geq 4$. In particular, $t(n,4) \leqslant 0.4932n$.

Twins in *random* words were not much studied before, although the proofs of the upper bounds in [3,6] were obtained via the probabilistic method. The only work addressing this issue directly is a recent paper by He et al. [13], where it is proved that, with high probability, binary words of length n contain twins of size $n/2 - \omega\sqrt{n}$, for any function $\omega = \omega(n)$ tending to infinity with n. Moreover, the authors of [13] formulate a striking conjecture that almost every binary word with even numbers of ones and zeros is a perfect pair of twins. If true, this would imply that, with high probability, a random binary word of an *even* length n contains twins of size at least $n/2 - 1$ (when the number of ones is odd, this is the best one can get).

1.2. Our Results

In this paper, we study twins in a *random word* $W_k(n)$ obtained by drawing one with probability k^{-n} out of all k-ary words of length n. Equivalently, one could toss a k-sided fair die, independently, n times. Either way, this is an equiprobable space.

Our main result improves the lower bound (1.1) for all but very few ternary ($k = 3$) words of length n.

Table 1. Comparing bound (1.2) of Bukh and Zhou [6] (for all words) with Theorem 1.2 (for almost all words)

k	3	4	5	10	50	100	200	400
$\left(\frac{k}{81}\right)^{1/3}$	0.333	0.367	0.395	0.498	0.851	1.073	1.352	1.703
$\frac{1.64k}{k+1}$	1.230	1.312	1.367	1.491	1.608	1.624	1.632	1.636

Theorem 1.1. *With probability* $1 - e^{-\Omega(n/(\log n))}$,

$$t(W_3(n)) \geq 0.411n.$$

The proof involves computer-assisted calculations (see Appendix A). It seems plausible that substantially stronger computational devices could bring further improvements of the bound. Moreover, by using classical methods (without computers) we provide a slightly worse bound, namely, $0.375n$ (cf. Example 3.8 in Sect. 3.4).

For larger alphabets we have the following result. We present it in a form which emphasizes the improvement over the deterministic bound (1.1). We say that an event $\mathcal{E}_n$ holds *asymptotically almost surely (a.a.s.)* if $\Pr(\mathcal{E}_n) \to 1$ as $n \to \infty$.

Theorem 1.2. *For each* $k \geq 3$ *and large* n, *a.a.s.*

$$t(W_k(n)) \geq \frac{1.64k}{k+1} \cdot \frac{n}{k},$$

This gives an improvement upon the (deterministic) estimates (1.1) and, for all $k \leqslant 354$, upon (1.2) (see Table 1).

The proof uses a tool called the Boosting Lemma (Lemma 3.4). It is stated in terms of a special model of random words which allows for iterative enhancement of the twin length, while adding a new letter to the alphabet. This new model assumes that the numbers of occurrences of letters are fixed in such a way that the model is asymptotically equivalent to $W_k(n)$.

1.3. Multiple Twins

We also consider a more general notion of *multiple* twins. By r-*twins* in a word w we mean r disjoint identical subwords of w. Let $t^{(r)}(w)$ be the maximum length of r-twins in w and let $t^{(r)}(n,k)$ be the minimum of $t^{(r)}(w)$ over all words w of length n from a k-letter alphabet. By the results of [3,6] we know, respectively, that $t^{(r)}(n,k) \sim n/r$ when $r \geqslant k$, and

$$t^{(r)}(n,k) \geqslant C_r \cdot k^{1/\binom{2r-1}{r}} \cdot \frac{n}{k} - O(1), \qquad (1.3)$$

for $k > r \geqslant 3$, where $C_r = \left(\frac{1}{2r-1}\right)^{1+1/\binom{2r-1}{r}}$ and the $O(1)$ term depends on r and k only. Estimate (1.3) was only mentioned in the concluding remarks in [6] and the explicit form of C_r was not given. However, based on the proof of inequality (1.2) therein, it is quite obvious how it should be computed.

TABLE 2. Comparing bound (1.3) of Bukh and Zhou [6] (for all words) with Theorem 1.3 (for almost all words)

(r,k)	$(3,4)$	$(3,10)$	$(3,100)$	$(3,1000)$	$(3,10^{10})$	$(4,10^{10})$	$(4,10^{40})$
$C_r \cdot k^{1/\binom{2r-1}{r}}$	0.196	0.214	1.036	0.340	1.703	0.261	1.878
$\Pi_{r,k}$	1.016	1.036	1.041	1.041	1.041	1.003	1.003

For a random word $W_k(n)$ we get the following estimates which, again, for small r and k yield better lower bounds than (1.3) (see Table 2). For $r,k \geq 2$, let $\Pi_{r,k} = \prod_{j=r+1}^{k} \frac{j^r}{j^r-1}$.

Theorem 1.3. *For every $k > r \geq 3$, a.a.s.*

$$t^{(r)}(W_k(n)) \geq \Pi_{r,k} \cdot \frac{n}{k} - o(n).$$

1.4. Organization

In the next section we prove Theorem 1.1. Section 3 begins with a short proof of the known bound $t^{(r)}(n,k) \sim n/r$, $r \geq k$, for random words. We do so for self-containment, as the proof in [3] (for *all* words) is quite involved. The next subsection contains a standard proof of an asymptotic equivalence between $W_k(n)$ and another model of random words. Then comes the crucial Boosting Lemma, while the last subsection of Sect. 3 brings applications of the Boosting Lemma, among them short proofs of Theorems 1.2 and 1.3, the former utilizing also Theorem 1.1. The last section contains some remarks and open problems, while the Appendix presents a Maple code used for derivation of the data collected in Table 3 as well as a proof of a technical estimate needed in Sect. 3.3.

2. Computer Assisted Bound

In this section, we prove Theorem 1.1.

Proof of Theorem 1.1. Fix a positive integer s and split the set of positions of a ternary random word $W_3(n)$ into $m := n/s$ segments of length s (we assume $s|n$). For $1 \leq j \leq m$ and $1 \leq t \leq \lfloor \frac{s}{2} \rfloor$, let X_t^j be the indicator random variable such that $X_t^j = 1$ if the j-th segment of a random word $W_3(n)$ contains twins of length t but not $t+1$; otherwise $X_t^j = 0$. Furthermore, define $X_t := \sum_{j=1}^{m} X_t^j$.

First we calculate the expected value of X_t. Let λ_t count the number of ternary words of length s with twins of length t but not $t+1$. Clearly, $\sum_{t=1}^{\lfloor s/2 \rfloor} \lambda_t = 3^s$. In general, finding (or even tightly approximating) λ_t does not seem to be an easy problem. However, for small s one can use a computer program to determine λ_s. In Table 3 we present all values of λ_t for all $6 \leq s \leq 14$ and $1 \leq t \leq \lfloor s/2 \rfloor$ (see Appendix A for more details). Since X_t has the binomial distribution $\mathrm{Bin}(m, \lambda_t/3^s)$, its expectation is

Table 3. The exact values of λ_t for all $6 \leq s \leq 14$ and $1 \leq t \leq \lfloor s/2 \rfloor$

s	6	7	8	9	10	11	12	13	14
λ_1	42	6	Zero						
λ_2	594	1086	822	288	42				
λ_3	93	1095	5118	11010	10806	5292	1350	162	
λ_4	Zero		621	8385	43776	106032	123750	75810	24894
λ_5					4425	65823	373638	992244	1312530
λ_6							32703	526107	3196644
λ_7									248901

$$\mathbb{E}(X_t) = \frac{\lambda_t}{3^s} m = \frac{\lambda_t}{s3^s} n.$$

Moreover, a standard application of the Chernoff inequality (see, e.g., ineq. (2.9) in [15]) together with the union bound yields that for all $1 \leq j \leq m$ and $1 \leq t \leq \lfloor \frac{s}{2} \rfloor$, X_t is highly concentrated around its mean, i.e.,

$$X_t = \frac{\lambda_t}{s3^s}(1 + o(1))n$$

with probability $1 - e^{-\Omega(n/(\log n))}$.

To finish the proof of Theorem 1.1 we construct twins in $W_3(n)$ as follows. Let, for each $1 \leq j \leq m$, A_j and B_j be a pair of the longest twins in the j-th segment. Observe that concatenating A_j and B_j over all m segments yields, a.a.s., twins $A_1 A_2 \dots A_m$ and $B_1 B_2 \dots B_m$ of length

$$\sum_{t=1}^{\lfloor s/2 \rfloor} tX_t = \sum_{t=1}^{\lfloor s/2 \rfloor} \frac{t\lambda_t}{s3^s}(1 + o(1))n = \varrho_s(1 + o(1))n,$$

where

$$\varrho_s := \sum_{t=1}^{\lfloor s/2 \rfloor} \frac{t\lambda_t}{s3^s}.$$

From the last column of Table 3 we compute that

$$\varrho_{14} = \frac{4 \cdot 24894 + 5 \cdot 1312530 + 6 \cdot 3196644 + 7 \cdot 248901}{14 \cdot 3^{14}} = \frac{27584397}{66961566} > 0.4119.$$

This completes the proof. □

3. General Bounds

3.1. More Twins than Letters

It was shown in [3] that for all $r \geq k$

$$t^{(r)}(n,k) \geq \frac{n}{r} - O\left(\left(\frac{\log\log n}{\log n}\right)^{1/4}\right).$$

For self-containment we insignificantly improve this lower bound for *almost* all words. The new bound holds with probability so close to 1 that a passage to the alternative model employed in Sect. 3.4 is still possible.

To this end, let us recall the main observation in [3]: if there is a partition of a k-word w of length n into m segments $w_1 w_2 \cdots w_m$ of sizes $N := n/m$ (we assume $m|n$) and in each segment each letter occurs at least μ times, then one can construct r-twins in w via a natural interlacing. Namely, Twin 1 consists of μ elements a_1 of w_1, μ elements a_2 of $w_2, \ldots, \mu$ elements a_k of w_k, followed by μ elements a_1 of w_{r+1}, μ elements a_2 of w_{r+2}, and so on, for as long as there are still at least $r-1$ segments ahead. Twin 2 follows the same pattern except that it begins with μ elements a_1 of w_2, that is, it is shifted by one segment with respect to Twin 1. Then Twin 3 begins with μ elements a_1 of w_3, etc.

This way all r twins are disjoint and they use together $k\mu$ elements from each segment except the first $k-1$ ones where the consumption gradually grows from μ to $(k-1)\mu$, and, in the worst case, when $m = r-1 \pmod r$, the last $r-1$ ones, where the consumption declines from $(k-1)\mu$ to μ, leaving the last $r-k$ segments completely untouched. The exact worst case count yields that together twins $T_1, \ldots, T_r$ cover at least

$$[m-(k-1)-(r-1)]\,k\mu + 2\binom{k}{2}\mu = (m-r+1)k\mu \qquad (3.1)$$

elements of w.

The R-H-S of (3.1) is, given $m \to \infty$ as $n \to \infty$, very close to $mk\mu$, which, in turn, will be close to n, provided $\mu \sim N/k$. To get these relations, the authors of [3] developed a regularity lemma for words. Here we are in a more comfortable situation as we are dealing with random words. Therefore, a simple application of Chernoff's bound yields the following result.

Lemma 3.1. *For $r \geq k \geqslant 2$, with probability at least $1 - O(n^{-k+1/3})$,*

$$t^{(r)}(W_k(n)) = n/r - O\left(n^{2/3}\sqrt{\log n}\right).$$

Proof. Split the set of positions of $W_k(n)$ into $m := n^{1/3}$ consecutive segments of length $N := n^{2/3}$. For $j = 1, \ldots, m$ and $i = 1, \ldots, k$, let $X := X_i^j$ be the number of elements a_i in the j-th segment. Then $\mathbb{E}X = N/k$ and, by Chernoff's bound from [15], ineq. (2.6), we have

$$\mathbb{P}(X \leq (1-\varepsilon)\mathbb{E}X) \leq n^{-k},$$

where $\varepsilon = kn^{-1/3}\sqrt{2\log n}$. As we only have $km = O(n^{1/3})$ random variables X_i^j, with probability at least $1 - O(n^{-k+1/3})$, they all satisfy the opposite bound, that is, for all $j = 1, \ldots, m$ and $i = 1, \ldots, k$,

$$X_i^j \geq \mu := (1-\varepsilon)n^{2/3}/k = n^{2/3}/k - O(n^{1/3}\sqrt{\log n}).$$

Upon applying (3.1) with this μ, we conclude that the number of elements uncovered by the r-twins is, indeed, at most $n-(m-r+1)k\mu = O(n^{2/3}\sqrt{\log n})$. □

Notice that for $r = k = 2$, the estimate for $t(W_2(n))$ in the lemma is slightly weaker than the one from [13] mentioned in the introduction. However, this fact does not affect our results, as all we really need in the proofs (see Example 3.8 and the proof of Theorem 1.3) is an estimate $t^{(r)}(W_k(n)) = n/r - o(1)$. Besides, it is not clear how to generalize the result from [13] to other values of k and r.

3.2. Equivalence of Models of Random Words

Guided by analogy to random graphs, we consider two basic models of random words: the binomial and the fixed-letter-count model. Given positive integers k and n, an alphabet $A = \{a_1, \ldots, a_k\}$, and constants $0 \leq p_1, \ldots, p_k \leq 1$, where $p_1 + \cdots + p_k = 1$, the *binomial random word* $W(n; p_1, \ldots, p_k)$ is a sequence of independent random variables $(X_1, \ldots, X_n)$, where for each $j = 1, \ldots, n$ and $i = 1, \ldots, k$ we have $\mathbb{P}(X_j = a_i) = p_i$. Here we are exclusively interested in the special, equiprobable instance $W_k(n) := W(n; 1/k, \ldots, 1/k)$.

As a technical tool rather than an object of genuine interest we also define another model of a random word. Given integers $0 \leq M_1, \ldots, M_k \leq n$, where $M_1 + \cdots + M_k = n$, *the fixed-letter-count random word* $W(n; M_1, \ldots, M_k)$ is obtained by taking uniformly at random a permutation (with repetitions) of n elements, among which, $i = 1, \ldots, k$, there are M_i elements a_i. Thus, in the latter model, we restrict ourselves to words with prescribed numbers of each letter and every such word has the same probability $\binom{n}{M_1,\ldots,M_k}^{-1}$ to be chosen.

The asymptotic equivalence between the two models goes smoothly one way (from fixed-letter-count to binomial), but to proceed the other way the models lack monotonicity, so we are forced to recourse to an analog of Pittel's inequality (see, e.g., ineq. (1.6) in [15]) which bring, however, some limitations.

Let Q be a property (subset) of words of length n over alphabet A. Although more general statements can be easily proved, we restrict ourselves to the special case of the model $W_k(n)$ and also to limiting probabilities equal to 1 only. The proofs follow those for random graphs (cf. Section 1.4 in [15], in particular, proofs of Prop. 1.12 and of Pittel's inequality (1.6) therein) and rely on the law of total probability and the fact that the space of $W_k(n)$ conditioned on M_i's being the numbers of occurrences of the elements $a_i \in A$, $i = 1, \ldots, k$, coincides with that of $W(n; M_1, \ldots, M_k)$.

Proposition 3.2. *If for all $M_1, \ldots, M_k$ such that $M_1 + \cdots + M_k = n$ and $M_i = n/k + O(\sqrt{n})$, $i = 1, \ldots, k$, $\mathbb{P}(W(n; M_1, \ldots, M_k) \in Q) \to 1$ as $n \to \infty$, then $\mathbb{P}(W_k(n) \in Q) \to 1$ as $n \to \infty$.*

Proof. Let C be a large constant and define (for each n)

$$\mathcal{M}_C = \{(M_1,\dots,M_k) : M_1+\cdots+M_k = n \quad \text{and} \quad |M_i - n/k| \le C\sqrt{n}\}.$$

Let $(M_1^*,\dots,M_k^*)$ minimize $\mathbb{P}(W(n;M_1,\dots,M_k)\in Q)$ over $\mathcal{M}_C$. Finally, let X_i be the number of occurrences of letter a_i in $W_k(n)$. Note that each X_i has binomial distribution with expectation n/k and variance less than n/k. Then, by the law of total probability

$$\mathbb{P}(W_k(n)\in Q) \ge \mathbb{P}(W(n;M_1^*,\dots,M_k^*)\in Q)\mathbb{P}((X_1,\dots,X_k)\in\mathcal{M}_C).$$

By assumption, $\mathbb{P}(W(n;M_1^*,\dots,M_k^*)\in Q)\to 1$ as $n\to\infty$. By Chebyshev's inequality applied together with the union bound,

$$\mathbb{P}((X_1,\dots,X_k)\notin\mathcal{M}_C) \le k\frac{n/k}{C^2 n} = C^{-2}.$$

Thus, $\liminf_{n\to\infty}\mathbb{P}(W_k(n)\in Q)\ge 1-C^{-2}$. As this is true for every C, letting $C\to\infty$ yields $\lim_{n\to\infty}\mathbb{P}(W_k(n)\in Q)=1$. □

Proposition 3.3. *If* $\mathbb{P}(W_k(n)\in Q) = 1-o(n^{-k/2})$ *as* $n\to\infty$, *then for all* $M_i = n/k+\omega_i$, *where* $|\omega_i|\le\sqrt{(n\log n)/(3k^2)}$ *and* $\sum_i\omega_i = 0$, $\mathbb{P}(W(n;M_1,\dots,M_k)\in Q)\to 1$ *as* $n\to\infty$.

Proof. Fix $M_1,\dots,M_k$ as in the statement of the proposition. By the law of total probability, we obtain

$$\begin{aligned}\mathbb{P}(W_k(n)\notin Q) &= \sum_{M_1',\dots,M_k'}\mathbb{P}(W(n;M_1',\dots,M_k')\notin Q)\binom{n}{M_1',\dots,M_k'}\frac{1}{k^n}\\ &\ge \mathbb{P}(W(n;M_1,\dots,M_k)\notin Q)\binom{n}{M_1,\dots,M_k}\frac{1}{k^n},\end{aligned}$$

from which we get

$$\mathbb{P}(W(n;M_1,\dots,M_k)\notin Q) \le \frac{k^n}{\binom{n}{M_1,\dots,M_k}}\mathbb{P}(W_k(n)\notin Q).$$

It remains to estimate the ratio $\frac{k^n}{\binom{n}{M_1,\dots,M_k}}$. Using Stirling's formula several times, we get

$$\begin{aligned}\frac{k^n}{\binom{n}{M_1,\dots,M_k}} &= O(n^{(k-1)/2})\prod_{i=1}^{k}(1+k\omega_i/n)^{M_i}\\ &= O(n^{(k-1)/2})\exp\left\{\sum_{i=1}^{k}k\omega_i M_i/n\right\}.\end{aligned}$$

Now since $M_i = n/k+\omega_i$ and $\sum_i\omega_i=0$, we get

$$\sum_{i=1}^{k}k\omega_i M_i/n = \sum_{i=1}^{k}k\omega_i\left(\frac{1}{k}+\frac{\omega_i}{n}\right) = \sum_{i=1}^{k}\frac{k\omega_i^2}{n}\le\frac{\log n}{3}$$

and so $k^n/\binom{n}{M_1,\dots,M_k} = O(n^{k/2-1/6})$, which yields that $\mathbb{P}(W_k(n)\notin Q)=o(1)$. □

3.3. Boosting Lemma

Fix $2 \le r \le k$ and observe that if for some $\lambda = \lambda(n) > 0$ and all $n \ge n_0$, we have $t^{(r)}(n,k) \ge \lambda n$, then also, by dropping the least frequent letter, $t^{(r)}(N,k+1) \ge \frac{\lambda k}{k+1}N$, provided $N \ge \frac{k+1}{k}n_0$. In this section, we show that for random words this trivial bound can be improved if one considers the fixed-letter-count model.

To get a similar result for the binomial model $W_k(n)$ one has to switch first to the fixed-letter-count model $W(n; M_1, \ldots, M_k)$ and then back to $W_{k+1}(N)$, the switches facilitated, respectively, by Propositions 3.3 and 3.2. The reason for switching is that the fixed-letter-count model can be broken into two phases allowing the enlargement of the twins.

Indeed, let $M_1, \ldots, M_{k+1}$ be given such that $\sum_{i=1}^{k+1} M_i = N$. We generate the random word $W(N; M_1, \ldots, M_{k+1})$ by first permuting all $n := M_1 + \cdots + M_k$ letters from $A \setminus \{a_{k+1}\}$ (Phase 1). This can be done in precisely $\binom{n}{M_1,\ldots,M_k}$ ways. Then we throw in the M_{k+1} letters a_{k+1} which can go anywhere between the previously distributed letters (Phase 2). This can be done, by the formula for the number of ways to allocate M_{k+1} balls into $n+1$ bins, in $\binom{N}{M_{k+1}}$ ways. Note that the product of these two numbers is, indeed, $\binom{N}{M_1,\ldots,M_{k+1}}$ and that the outcome of Phase 1 is precisely the random word $W(n; M_1, \ldots, M_k)$.

Lemma 3.4. (Boosting Lemma) *For $2 \le r \le k$ and n sufficiently large, let a partition $n = M_1 + \cdots + M_k$ into nonnegative integers be given. If $M_{k+1} = n/k + O(\sqrt{n})$ and, for some $\lambda = \lambda(n) > 0$, a.a.s. $t^{(r)}(W(n; M_1, \ldots, M_k)) \ge \lambda n$, then a.a.s.*

$$t^{(r)}(W(N; M_1, \ldots, M_k, M_{k+1})) \ge \left(1 + \frac{1}{(k+1)^r - 1}\right)\frac{\lambda k}{k+1}N(1 - o(1)),$$

where $N = n + M_{k+1}$.

Note that the factor of $\frac{\lambda k}{k+1}N$ comes for free already after Phase 1, so that the actual improvement sits in the parentheses. Also, notice that

$$\frac{n}{N} = \frac{k}{k+1}\left(1 + O\left(\frac{1}{\sqrt{n}}\right)\right).$$

In the proof of Lemma 3.4 we will need a technical estimate on a ratio of binomial coefficients, the proof of which is deferred to Appendix B.

Fact 3.5. *Let $M = \Theta(N)$ and $\ell^2 = o(N)$. Then,*

$$\frac{\binom{N-\ell}{M-\ell}}{\binom{N}{M}} = \left(\frac{M}{N}\right)^{\ell}\left(1 + O\left(\frac{\ell^2}{N}\right)\right).$$

Proof of Lemma 3.4. We generate $W(N; M_1, \ldots, M_{k+1})$ in two phases as described prior to the statement of the lemma. Let W' be the outcome of Phase 1, that is, an instance of $W(n; M_1, \ldots, M_k)$. Further, let Q' be the event that $t^{(r)}(W(n; M_1, \ldots, M_k)) \ge \lambda n$ and let Q be the ultimate event that

$$t^{(r)}(W(N; M_1, \ldots, M_k, M_{k+1})) \ge \left(1 + \frac{1}{(k+1)^r - 1}\right)\frac{\lambda k}{k+1}N(1 - o(1)).$$

By the law of total probability,

$$\mathbb{P}(Q) = \sum_{W'} \mathbb{P}(Q|W')\mathbb{P}(W') = \sum_{W' \in Q'} \mathbb{P}(Q|W')\mathbb{P}(W') + o(1). \tag{3.2}$$

We now focus on $\mathbb{P}(Q|W')$ with $W' \in Q'$, that is, fixing an instance W' of Phase 1 with $t^{(r)}(W) \geq \lambda n$, we are going to thoroughly investigate the outcome of Phase 2. Fix $W' = w_1 \cdots w_n$ and r-twins $T_1, \ldots, T_r$ therein of length $|T_1| = \cdots = |T_r| = \lambda n$ (we may assume that λn is an integer). We treat the $n+1$ spaces before, between, and after the letters of W' as bins and the M_{k+1} letters a_{k+1} as balls. Precisely, bin b_0 is in front of w_1, for $i = 1, \ldots, n-1$, bin b_i lies between w_i and w_{i+1}, and bin b_n is to the right of w_n.

We group the bins lying immediately to the right of the elements of the r-twins $T_1, \ldots, T_r$ into *r-tuples of bins* denoted by $R_1, \ldots, R_{\lambda n}$. Formally, if $T_j = w^{(j)}_{i_1} \cdots w^{(j)}_{i_{\lambda n}}$, $j = 1, \ldots, r$, then the r-tuple of bins R_ℓ consists of the bins $b^{(1)}_{i_\ell}, \ldots, b^{(r)}_{i_\ell}$, for $\ell = 1, \ldots, \lambda n$.

Example 3.6. Let $k = r = 3$, $n = 27$, and

$$W' = w_1 \cdots w_{27} = b\ a\ a\ c\ a\ c\ b\ b\ a\ b\ c\ a\ c\ a\ c\ a\ b\ b\ c\ b\ a\ c\ c\ c\ b\ a\ b.$$

There are 3-twins of length 5 here, each forming the word *aabcc*, namely, $T_1 = w_2w_3w_8w_{11}w_{19}$, $T_2 = w_5w_9w_{10}w_{13}w_{23}$, and $T_3 = w_{12}w_{16}w_{20}w_{22}w_{24}$. For instance, consider the first triple of bins, $R_1 = \{b_2, b_5, b_{12}\}$. If the letter d is inserted into each of these three bins, a longer word

$$b\ a\ \underline{d}\ a\ c\ a\ \underline{d}\ c\ b\ b\ a\ b\ c\ a\ \underline{d}\ c\ a\ c\ a\ b\ b\ c\ b\ a\ c\ c\ c\ b\ a\ b,$$

is obtained, and the length of the twins, each forming now the word *adabcc*, increases by one.

Clearly, if each bin from an r-tuple receives s balls (read: s letters a_{k+1}), then the length of the twins can be extended by s. Since we would like to utilize several r-tuples of bins simultaneously, we categorize them with respect to the minimum number of balls.

For $1 \leq s \leq \log n$, let X_s be the number of r-tuples of bins with at least s balls in each bin and exactly s balls in some bin (i.e. there is a bin with exactly s balls). We call each such r-tuple of bins an *s-provider.* For each $\ell = 1, \ldots, \lambda n$, let I_ℓ be the indicator random variable such that $I_\ell = 1$ if R_ℓ is an s-provider and 0 otherwise. Hence, $X_s = \sum_{\ell=1}^{\lambda n} I_\ell$.

The event $\{I_\ell = 1\}$ is the set difference of two events: that each bin in R_ℓ has at least s balls and that each bin in R_ℓ has at least $s+1$ balls. Thus, setting $M := M_{k+1}$,

$$\mathbb{P}(I_i = 1) = \frac{\binom{N-rs}{M-rs}}{\binom{N}{M}} - \frac{\binom{N-rs-r}{M-rs-r}}{\binom{N}{M}}$$

and Fact 3.5 yields

$$\mathbb{P}(I_i = 1) = \left(\frac{M}{N}\right)^{rs} \left(1 - \left(\frac{M}{N}\right)^{r}\right) \left(1 + O\left(\frac{s^2}{N}\right)\right)$$

$$= \left(\frac{1}{k+1}\right)^{rs}\left(1-\left(\frac{1}{k+1}\right)^{r}\right)\left(1+O\left(\frac{1}{\sqrt{n}}\right)\right).$$

Thus, setting $\kappa = (k+1)^{-r}$, we have $\mathbb{E}X_s = \lambda n(1-\kappa)\kappa^s\left(1+O\left(1/\sqrt{n}\right)\right)$.

Our goal is to show a.a.s. simultaneous concentration of each X_s, $s \le \log n$, near its expectation. We are going to use Chebyshev's inequality (followed by the union bound over all s)

$$\mathbb{P}(|X_s - \mathbb{E}X_s| \ge \gamma \mathbb{E}X_s) \le \frac{VarX_s}{(\gamma \mathbb{E}X_s)^2}$$

with $\gamma = 1/\log n$. To facilitate the future use of the union bound, we need to show that $VarX_s = o((\mathbb{E}X_s)^2/\log^3 n)$, which will imply that $\frac{VarX_s}{(\gamma\mathbb{E}X_s)^2} = o(1/\log n)$. To this end we write

$$VarX_s = \mathbb{E}(X_s(X_s-1)) + \mathbb{E}X_s - (\mathbb{E}X_s)^2.$$

Note that $\{I_{\ell_1} = I_{\ell_2} = 1\} = A \setminus (B_1 \cup B_2)$, where A is the event that all $2r$ bins in R_{ℓ_1} and R_{ℓ_2} each contains at least s balls, while B_i, $i = 1, 2$, is the event that each bin in R_{ℓ_i} contains at least s balls and each bin in $R_{\ell_{3-i}}$ contains at least $s+1$ balls. Thus,

$$\begin{aligned}\mathbb{P}(I_{\ell_1} = I_{\ell_2} = 1) &= \mathbb{P}(A) - \mathbb{P}(B_1) - \mathbb{P}(B_2) + \mathbb{P}(B_1 \cap B_2)\\ &= \frac{\binom{N-2rs}{M-2rs}}{\binom{N}{M}} - 2\frac{\binom{N-2rs-r}{M-2rs-r}}{\binom{N}{M}} + \frac{\binom{N-2rs-2r}{M-2rs-2r}}{\binom{N}{M}}\end{aligned}$$

which by Fact 3.5 is equal to

$$\left(\kappa^{2s} - 2\kappa^{2s+1} + \kappa^{2s+2}\right)\left(1+O\left(\frac{1}{\sqrt{n}}\right)\right) = ((1-\kappa)\kappa^s)^2\left(1+O\left(\frac{1}{\sqrt{n}}\right)\right).$$

Thus,

$$\begin{aligned}VarX_s &= \lambda n(\lambda n - 1)\left((1-\kappa)\kappa^s\right)^2\left(1+O\left(\frac{1}{\sqrt{n}}\right)\right)\\ &\qquad + \lambda n(1-\kappa)\kappa^s\left(1+O\left(\frac{1}{\sqrt{n}}\right)\right) - \left(\lambda n(1-\kappa)\kappa^s\left(1+O\left(\frac{1}{\sqrt{n}}\right)\right)\right)^2\\ &= \left(\lambda n(1-\kappa)\kappa^s\right)^2 + O(n^{3/2})\\ &\qquad + O(n) - \left(\lambda n(1-\kappa)\kappa^s\right)^2 - O(n^{3/2}) = O(n^{3/2}) = o((\mathbb{E}X_s)^2/\log^3 n),\end{aligned}$$

with a big margin.

Hence,

$$\sum_{s=1}^{\lfloor \log n \rfloor} \mathbb{P}(|X_s - \mathbb{E}X_s| \ge \gamma\mathbb{E}X_s) = o(1),$$

and, in particular, a.a.s., for all $1 \le s \le \log n$,

$$\begin{aligned}X_s \ge \mathbb{E}X_s\left(1-\frac{1}{\log n}\right) &= \lambda n(1-\kappa)\kappa^s\left(1+O\left(\frac{1}{\sqrt{n}}\right)\right)\left(1-\frac{1}{\log n}\right)\\ &= \lambda n(1-\kappa)\kappa^s\left(1-O\left(\frac{1}{\log n}\right)\right).\end{aligned}$$

As each X_s contributes s towards an enlargement of the twins $T_1,\ldots,T_r$, we need to calculate $\sum_{s=1}^{\lfloor \log n\rfloor} sX_s$. Since for every positive integer p

$$\sum_{s=1}^{p} s\kappa^s = \sum_{s=1}^{\infty} s\kappa^s - \sum_{s=p+1}^{\infty} s\kappa^s = \frac{\kappa}{(1-\kappa)^2} - \frac{(1+(1-k)p)\kappa^{p+1}}{(1-\kappa)^2},$$

we get

$$\begin{aligned}\sum_{s=1}^{\lfloor \log n\rfloor} sX_s &\geq \lambda n(1-\kappa)\left(1-O\left(\frac{1}{\log n}\right)\right)\sum_{s=1}^{\lfloor \log n\rfloor} s\kappa^s\\ &= \lambda n(1-\kappa)\left(1-O\left(\frac{1}{\log n}\right)\right)\left(\frac{\kappa}{(1-\kappa)^2} - \Theta(\kappa^{\log n})\right)\\ &= \frac{\lambda\kappa}{1-\kappa}n(1-o(1)) = \frac{\lambda}{(1+k)^r-1}n(1-o(1)).\end{aligned}$$

Hence, still conditioning on W', a.a.s., there are in $W(N;M_1,\ldots,M_k,M_{k+1})$ twins of length at least

$$\lambda n + \frac{\lambda}{(1+k)^r-1}n(1-o(1)) = \left(1+\frac{1}{(1+k)^r-1}\right)\lambda n(1-o(1)).$$

As $n = \frac{k}{k+1}N(1+o(1))$, this means that $\mathbb{P}(Q|W') = 1-o(1)$ uniformly over all $W'\in Q'$, and, by (3.2), the lemma follows. □

Remark 3.7. By considering only one random variable X counting r-tuples of bins with no bin empty, the proof becomes a bit simpler, but the result is slightly weaker: a.a.s.

$$t^{(r)}(W(N;M_1,\ldots,M_k,M_{k+1})) \geq \left(1+\frac{1}{(k+1)^r}\right)\frac{\lambda k}{k+1}N.$$

We get rid of $o(1)$, but lose -1 in the denominator, so overall the bound is lower, though for large k the difference is insignificant.

3.4. Applications

In this subsection, we present applications of the Boosting Lemma (Lemma 3.4), most importantly, in the proofs of Theorems 1.2 and 1.3. Each time the scenario is the same: we pick an existing lower bound on the length of twins in $W_{k'}(n)$ (e.g., the bound in Theorem 1.1 above or any deterministic result from [3] or [6]), translate it via Proposition 3.3 to the fixed-letter-count model and then apply (iteratively) the Boosting Lemma to get a bound, still in the fixed-letter-count model, for the targeted value of $k > k'$. At the end we go back to the binomial model $W_k(n)$ via Proposition 3.2. For this we need to have our estimate valid for all M_i, $i=1,\ldots,k$, satisfying the assumptions of Proposition 3.2. This means, however, that also the input bound has to be valid for a corresponding range of M_i, $i=1,\ldots,k'$.

Let us now analyze what it really means. Fix k and let $M_i = n/k + w_i$, $i=1,\ldots,k$, where $|w_i| = O(\sqrt{n})$ and $\sum_{i=1}^{k} w_i = 0$. How these assumptions alter when we drop M_k? Let $n_{k-1} = M_1 + \cdots + M_{k-1}$. Then, for each $i = 1, \ldots, k-1$,

$$n_{k-1} = \frac{k-1}{k} n + \sum_{i=1}^{k-1} w_i = (k-1)M_i - (k-1)w_i - w_k,$$

thus

$$M_i = \frac{n_{k-1}}{k-1} + w_i + \frac{w_k}{k-1} = \frac{n_{k-1}}{k-1} + w_i^{(k-1)},$$

where $w_i^{(k-1)} = w_i + \frac{w_k}{k-1}$. Note that $\sum_{i=1}^{k-1} w_i^{(k-1)} = 0$ as it should. We may iterate this relation all the way down to k' (where we want to begin the process of applying Lemma 3.4), obtaining for $j = k-1, \ldots, k'$ and $i = 1, \ldots, k'$, with $n_j = M_1 + \cdots + M_j$,

$$M_i = \frac{n_j}{k} + w_i^{(j)}, \tag{3.3}$$

where

$$w_i^{(j)} = w_i + \frac{1}{j} \sum_{q=j+1}^{k} w_q. \tag{3.4}$$

Observe that $w_i^{(k')} = O(\sqrt{n})$, so we stay within the range required in Proposition 3.3. Below we illustrate how to apply the Boosting Lemma.

Example 3.8. Let $r = 2$ and $k = 3$. We know, either by our Lemma 3.1 or by the result in [3], that a.a.s. a random binary word $W_2(n)$ contains twins of length $\left(\frac{1}{2} - o(1)\right) n$. We would like to deduce from this, via Lemma 3.4, a lower bound on the length of twins in the random ternary word $W_3(n)$ (so, we take $k' = 2$ here).

Fix $M_i = n/3 + w_i$, $i = 1, 2, 3$, where $|w_i| = O(\sqrt{n})$ and $w_1 + w_2 + w_3 = 0$. Suppressing M_3, we get, with $n' = n - M_3$, an instance of $W(n'; M_1, M_2)$ which satisfies the assumptions of Proposition 3.3, that is, for $i = 1, 2$, we have $M_i = n'/2 + w_i^{(2)}$, where $w_i^{(2)} = w_i + \frac{w_3}{2} = O(\sqrt{n})$. Thus, we may conclude that a.a.s. $t^{(2)}(W(n'; M_1, M_2)) \geq \lambda n'$ with $\lambda = \frac{1}{2} - o(1)$. In turn, by Lemma 3.4, a.a.s.

$$t^{(2)}(W(n; M_1, M_2, M_3)) \geq \left(1 + \frac{1}{8}\right) \frac{2\lambda}{3} n(1 - o(1)) \geq 0.375n(1 - o(1)).$$

Since this is true for all choices of M_i as above, by Proposition 3.2, we finally get that a.a.s. $t^{(2)}(W_3(n)) \geq 0.375n(1 - o(1))$. This is much less than the bound in Theorem 1.1, so the result has some value only for computer-skeptical readers. On the other hand, it is still better than the bound $t^{(2)}(W_3(n)) \geq 0.34n(1 - o(1))$ in (1.1), though the latter holds for *all* ternary words.

Continuing with this example, let us iterate applications of Lemma 3.4 till, say, $k = 10$. Skipping details, we see that the obtained bound is a.a.s.

$$t^{(2)}(W_{10}(n)) \geq \frac{9}{8} \cdot \frac{16}{15} \cdot \frac{25}{24} \cdot \frac{36}{35} \cdot \frac{49}{48} \cdot \frac{64}{63} \cdot \frac{81}{80} \cdot \frac{100}{99} \cdot \frac{n}{10}(1 + o(1))$$

$$= \frac{15}{11} \cdot \frac{n}{10}(1 - o(1)) = 1.\overline{36} \cdot \frac{n}{10}(1 - o(1)).$$

It is perhaps interesting to compare the above bound with $1.337(n/10)$ – one obtained by the simpler and weaker version of Lemma 3.4 mentioned in Remark 3.7.

Proof of Theorem 1.2. As a starting point we take our computer-assisted Theorem 1.1, so we set $k' = 3$. Preparing for the final transition to the binomial model, fix any $M_i = n/k + w_i$, $i = 1, \ldots, k$, where $|w_i| = O(\sqrt{n})$ and $\sum_{i=1}^{k} w_i = 0$. By (3.3) with $j = 3$, we then have $M_i = n_3/3 + w_i^{(3)}$, where $n_3 = M_1 + M_2 + M_3$ and $w_i^{(3)} = O(\sqrt{n})$, $i = 1, 2, 3$, by (3.4).

By Theorem 1.1, $t(W_3(n)) \geq 0.411n$ with probability at least $1 - e^{-\Omega(n/(\log n))} = 1 - o(n^{-k/2})$. Thus, by Proposition 3.3, a.a.s. $t(W(n_3; M_1, M_2, M_3)) \geq 0.411n$, and, in turn, by Boosting Lemma (Lemma 3.4) with $\lambda = 0.411$, a.a.s. $t(W(n_4; M_1, M_2, M_3, M_4)) \geq \frac{16}{15} \cdot 1.233 \cdot \frac{n_4}{4}$. We iterate this transition (or use induction) until reaching the random word $W(n; M_1, \ldots, M_k)$. Observe that

$$\prod_{j=4}^{k} \frac{j^2}{j^2 - 1} = \prod_{j=4}^{k} \frac{j^2}{(j-1)(j+1)} = \frac{4^2}{3 \cdot 5} \cdot \frac{5^2}{4 \cdot 6} \cdot \frac{6^2}{5 \cdot 7} \cdots \frac{k^2}{(k-1)(k+1)} = \frac{4k}{3(k+1)}.$$

Hence, we get, a.a.s.,

$$t^{(2)}(W(n; M_1, \ldots, M_k)) \geq \prod_{j=4}^{k} \frac{j^2}{j^2 - 1} \cdot (3\varrho_{14}) \cdot \frac{n}{k}(1 - o(1)) \geq \frac{1.64k}{k+1} \cdot \frac{n}{k}.$$

(We drop 0.004 to "kill" the error term $1 - o(1)$.)

As this is true for all choices of $M_i = n/k + w_i$, we are in position to apply Proposition 3.2 and conclude that a.a.s. $t^{(2)}(W_k(n)) \geq \frac{1.64k}{k+1} \cdot \frac{n}{k}$. □

Proof of Theorem 1.3. This proof is very similar except that we now take $k' = r$ and launch off with the bound $t^{(r)}(W_{k'}(n)) = n/r - o(n)$ from Lemma 3.1 (or its deterministic counterpart from [3]). Other minor differences are that for $r \geq 3$ there is no nice formula for the product $\Pi_{r,k} = \prod_{j=r+1}^{k} \frac{j^r}{j^r - 1}$ and, unlike before, we cannot get rid of the error term $1 - o(1)$. We omit the details. □

4. Concluding Remarks

Let us conclude the paper with some open questions and suggestions for future studies. The first problem naturally concerns twins in ternary words.

Problem 4.1. Does $t(n, 3) = n/2 - o(n)$?

Even if the answer is negative it can still be true that *almost all* ternary words contain twins of such length, that is, a.a.s. $t(W_3(n)) = n/2 - o(n)$. It is known, as demonstrated in [6], that a.a.s. $t(W_4(n)) \leqslant 0.4932n$ which makes a similar statement for random *quaternary* words false. In general, it is not clear how close $t(W_k(n))$ and $t(n, k)$ are from each other.

Problem 4.2. Does $t(W_k(n)) = t(n,k) + o(n)$ hold a.a.s. for all $k \geqslant 2$?

By the result of Axenovich, Person, and Puzynina [3] we know that this is the case for $k = 2$.

One may also consider more restricted notions of twins in words reflecting their placement in the word. For instance, it may happen that twins occupy together a connected segment forming a *shuffle square*, as in the example below:

$$a\; b\; \boxed{c}\; \boxed{a}\; \boxed{c}\; \boxed{b}\; \boxed{a}\; \boxed{b}\; c\; b\; a\; c.$$

As proved recently by Bulteau, Jugé, and Vialette [4], there exist arbitrarily long words over a 6-letter alphabet containing no shuffle squares. It is not known, however, if the size of the alphabet in this result is optimal and, more generally, what is the maximum length of a shuffle square guaranteed to be present in *every* (long) k-letter word, $k = 2, \ldots, 5$.

Here, we formulate this question for random words.

Problem 4.3. What is the expected maximum length of a shuffle square in a random word $W_k(n)$?

This question sounds particularly intriguing in the light of a recent conjecture by He, Huang, Nam, and Thaper [13] (mentioned already in Sect. 1.1), that almost every binary word (with even numbers of ones and zeros) is a shuffle square. For more on counting shuffle squares and their variants see, e.g., [14]. Similar problems can also be considered for *shuffle cubes*, or more generally, for arbitrary *shuffle r-powers*, in analogy to general r-twins.

An even more restricted version of twins is obtained when each twin in a shuffle square occupies itself a connected segment, like in the example below:

$$a\; b\; \boxed{c}\; \boxed{a}\; \boxed{b}\; \boxed{c}\; \boxed{a}\; \boxed{b}\; c\; b\; a\; c.$$

This basic structure is well-known and widely studied in combinatorics on words under the name of a *square* or a *repetition* (see [17,18]). By the famous result of Thue [19] we know that there exist ternary words of any length with no squares altogether. In the binary case, it is known that there exist arbitrarily long words avoiding squares of length greater than 2, as proved by Fraenkel and Simpson [11]. However, not much is known about squares in random words.

Problem 4.4. What is the expected maximum length of a square in a random word $W_k(n)$?

Declarations

Conflict of Interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Appendix A: Maple Code

Here we present a simple program written in *Maple* that counts the number of ternary words of length s that contain twins of length at least t. This easily allows one to calculate the number of ternary words of length s with twins of length t but not $t+1$ (defined as λ_t in the proof of Theorem 1.1). Indeed, λ_t is equal to the difference between the number of words having twins of length at least t minus the one with twins of length at least $t+1$.

For the sake of clarity, we did not attempt to optimize our program.

```
with(combinat): # Load combinatorial functions

##########################################################
# This procedure returns the number of ternary words
# of length s that contain twins of length at least t
##########################################################
count_twins := proc(s, t)
    local P_list, P, sizeP, K, sizeK, L, sizeL, L1, L2, counter, p, k,
     l, i:

    # Generate all ternary words of length s
    P_list := [seq(1, i = 1 .. s), seq(2, i = 1 .. s), seq(3, i = 1 ..
     s)]:
    P := permute(P_list, s):
    sizeP := nops(P):

    # Generate all 2t−subsets of {1,...,s}
    K := choose([seq(i, i = 1 .. seq_length)], 2*t):
    sizeK := nops(K):

    # Partition {1,...,2t} into two sets, each of size t
    L := setpartition([seq(i, i = 1 .. 2*t)], t):
    sizeL := nops(L):

    counter := 0: # Number of words with twins of length at least t

    for p to sizeP # Over all ternary words
    do
        for k to sizeK # Over all 2t subsets of {1,...,s}
        do
            for l to sizeL # Over all partitions of {1,...,2t}
            do
                L1 := L[l][1]: # Indices for the first possible twin
                L2 := L[l][2]: # Indices for the second possible twin

                if evalb(P[p][K[k][L1]] = P[p][K[k][L2]]) # Twins?
                then
                    # Twins found
                    counter := counter + 1:

                    # Move to the next word in P
                    l := sizeL: # Break the loop over l
                    k := sizeK: # Break the loop over k
                end if:
            end do: # Over l
        end do: # Over k
    end do: # Over p
    counter: # Return the value of the variable counter
end proc:
```

Listing 1. The main procedure.

In Listing 2 we show how the procedure from Listing 1 can be used yielding results summarized in Table 3.

```
seq_length := 6: # Length of the ternary words
min_twin_length := 1: # Min length of possible twins
max_twin_length := floor(seq_length/2): # Max length of possible twins

for i from max_twin_length by -1 to min_twin_length do
    c[i] := count_twins(seq_length, i): # Number of twins of length >=
     i
    if i = max_twin_length
    then
        d[i] := c[i]: # Since there are no twins of length
    max_twin_length+1, c[max_twin_length] stores the number of words
    with the longest twins of length max_twin_length
    else
        d[i] := c[i] - c[i + 1]:
    end if:
    printf("Number of words with the longest twins of length %d = %d\n
    ", i, d[i]):
end do:
```

Listing 2. Calling the main procedure.

Appendix B: Proof of Fact 3.5

Proof of Fact 3.5. We are going to use the following form of Stirling's formula:

$$N! = \sqrt{2\pi N}\left(\frac{N}{e}\right)^N \left(1 + O\left(\frac{1}{N}\right)\right).$$

Hence,

$$\frac{N!}{(N-\ell)!} = \sqrt{\frac{N}{N-\ell}} \cdot \frac{N^N}{(N-\ell)^{N-\ell}} \cdot e^{-\ell} \left(1 + O\left(\frac{1}{N}\right)\right).$$

Since

$$\sqrt{\frac{N}{N-\ell}} = \left(1+\frac{\ell}{N-\ell}\right)^{1/2} = 1+O\left(\frac{\ell}{N}\right),$$

we get

$$\frac{N!}{(N-\ell)!} = \frac{N^N}{(N-\ell)^{N-\ell}} \cdot e^{-\ell}\left(1+O\left(\frac{\ell}{N}\right)\right).$$

Consequently, as

$$\frac{\binom{N-\ell}{M-\ell}}{\binom{N}{M}} = \frac{M!/(M-\ell)!}{N!/(N-\ell)!},$$

we obtain

$$\frac{\binom{N-\ell}{M-\ell}}{\binom{N}{M}} = \frac{M^M/(M-\ell)^{M-\ell}}{N^N/(N-\ell)^{N-\ell}}\left(1+O\left(\frac{\ell}{N}\right)\right).$$

Further, since $1+x = e^{x+O(x^2)}$ whenever $x \to 0$,

$$\frac{M^M/(M-\ell)^{M-\ell}}{N^N/(N-\ell)^{N-\ell}} = \left(\frac{M}{N}\right)^{\ell} \frac{\left(1-\frac{\ell}{N}\right)^{N-\ell}}{\left(1-\frac{\ell}{M}\right)^{M-\ell}} = \left(\frac{M}{N}\right)^{\ell} \frac{e^{-\ell(N-\ell)/N+O(\ell^2/N)}}{e^{-\ell(M-\ell)/M+O(\ell^2/M)}}$$
$$= \left(\frac{M}{N}\right)^{\ell} e^{O(\ell^2/N)},$$

where the last equality follows, because $-\ell(N-\ell)/N+\ell(M-\ell)/M = \ell^2(M-N)/(MN)$ and $M = \Theta(N)$. Finally, as by assumption $\ell^2 = o(N)$, we have $e^{O(\ell^2/N)} = 1+O(\ell^2/N)$ and so

$$\frac{\binom{N-\ell}{M-\ell}}{\binom{N}{M}} = \left(\frac{M}{N}\right)^{\ell}\left(1+O\left(\frac{\ell^2}{N}\right)\right)\left(1+O\left(\frac{\ell}{N}\right)\right) = \left(\frac{M}{N}\right)^{\ell}\left(1+O\left(\frac{\ell^2}{N}\right)\right),$$

as required. □

References

[1] N. Alon, Y. Caro, I. Krasikov, Bisection of trees and sequences, in Combinatorics and Algorithms (Jerusalem, 1988). Discrete Mathematics, vol. 114, no. 1–3 (Amsterdam, North-Holland, 1993), pp. 3–7.

[2] M. Axenovich, Repetitions in graphs and sequences, in Recent Trends in Combinatorics, Springer 2016.

[3] M. Axenovich, Y. Person, and S. Puzynina, A regularity lemma and twins in words, Journal of Combinatorial Theory. Series A 120 (2013), 733–743.

[4] L. Bulteau, V. Jugé, and S. Vialette, On shuffled-square-free words, Theoretical Comp. Sci. 941 (2023), 91–103.

[5] B. Bukh and O. Rudenko, Order-isomorphic twins in permutations, SIAM J. Discrete Math. 34 (2020) 1620–1622.

[6] B. Bukh and L. Zhou, Twins in words and long common subsequences in permutations, Israel Journal of Mathematics 213 (2016), 183–209.

[7] F. R. K. Chung, P. Erdős, R. L. Graham, S. M. Ulam, and F. F. Yao, Minimal decompositions of two graphs into pairwise isomorphic subgraphs. Congr. Numer. 23 (1979) 3–18.

[8] A. Dudek, J. Grytczuk, and A. Ruciński, Variations on twins in permutations, Electronic Journal of Combinatorics 28 (2021), no. 3, P3.19.

[9] A. Dudek, J. Grytczuk, and A. Ruciński, On weak twins and up-and-down subpermutations, Integers 21A (2021), Ron Graham Memorial Volume, Paper No. A10, 17 pp.

[10] A. Dudek, J. Grytczuk and A. Ruciński, Tight multiple twins in permutations, Annals of Combinatorics, 25 (2021), 1075–1094.

[11] A. S. Fraenkel and R. J. Simpson, How Many Squares Must a Binary Sequence Contain?, Electronic Journal of Combinatorics 2 (1995) R2.

[12] H. Hàn, M. Kiwi, and M. Pavez-Signé, Quasi-random words and limits of word sequences, 2003.03664 [math.CO].

[13] X. He, E. Huang, I. Nam, and R. Thaper, Shuffle squares and reverse shuffle squares, 2109.12455 [math.CO].

[14] D. Henshall, N. Rampersad, and J. Shallit, Shuffling and Unshuffling, Bull. EATCS 107 (2012) 131–142.

[15] S. Janson, T. Łuczak, and A. Ruciński, *Random Graphs*, John Wiley and Sons, 2000.

[16] C. Lee, P. Loh, and B. Sudakov, Self-similarity of graphs. SIAM J. Discret. Math. 27(2) (2013) 959–972.

[17] M. Lothaire, Combinatorics on Words, Addison-Wesley, Reading, MA, 1983.

[18] N. Rampersad and J. Shallit, Repetitions in words, in V. Berthé and M. Rigo, eds., Combinatorics, Words and Symbolic Dynamics, Encyclopedia of Mathematics and Its Applications, vol. 159, Cambridge University Press, 2016, pp. 101–150.

[19] A. Thue, Über unendliche Zeichenreichen, *Norske vid. Selsk. Skr. Mat. Nat. Kl.* **7** (1906), 1–22. Reprinted in *Selected Mathematical Papers of Axel Thue*, T. Nagell, editor, Universitetsforlaget, Oslo, 1977, pp. 139–158.

Andrzej Dudek
Department of Mathematics
Western Michigan University
Kalamazoo MI
USA
e-mail: andrzej.dudek@wmich.edu

Jarosław Grytczuk
Faculty of Mathematics and Information Science
Warsaw University of Technology
Warsaw
Poland
e-mail: j.grytczuk@mini.pw.edu.pl

Andrzej Ruciński
Department of Discrete Mathematics
Adam Mickiewicz University
Poznan
Poland
e-mail: rucinski@amu.edu.pl

Received: 14 June 2022.
Accepted: 30 April 2023.