Advances in Applied Mathematics 80 (2016) 48–61

Contents lists available at ScienceDirect

Advances in Applied Mathematics

www.elsevier.com/locate/yaama

Perfect necklaces

Nicolás Alvarez [a], Verónica Becher [b,*], Pablo A. Ferrari [b], Sergio A. Yuhjtman [c]

[a] *Universidad Nacional del Sur, Argentina*
[b] *Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires & CONICET, Argentina*
[c] *Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Argentina*

ARTICLE INFO

Article history:
Received 20 January 2016
Received in revised form 23 April 2016
Accepted 6 May 2016
Available online xxxx

MSC:
68R15
05C38
05C45

Keywords:
Combinatorics on words
Necklaces
de Bruijn words
Statistical tests of finite size

ABSTRACT

We introduce a variant of de Bruijn words that we call perfect necklaces. Fix a finite alphabet. Recall that a word is a finite sequence of symbols in the alphabet and a circular word, or necklace, is the equivalence class of a word under rotations. For positive integers k and n, we call a necklace (k, n)-perfect if each word of length k occurs exactly n times at positions which are different modulo n for any convention on the starting point. We call a necklace perfect if it is (k, k)-perfect for some k. We prove that every arithmetic sequence with difference coprime with the alphabet size induces a perfect necklace. In particular, the concatenation of all words of the same length in lexicographic order yields a perfect necklace. For each k and n, we give a closed formula for the number of (k, n)-perfect necklaces. Finally, we prove that every infinite periodic sequence whose period coincides with some (k, n)-perfect necklace for some k and some n, passes all statistical tests of size up to k, but not all larger tests. This last theorem motivated this work.

* Corresponding author.
E-mail addresses: naa@cs.uns.edu.ar (N. Alvarez), vbecher@dc.uba.ar (V. Becher), pferrari@dm.uba.ar (P.A. Ferrari), syuhjtma@dm.uba.ar (S.A. Yuhjtman).

http://dx.doi.org/10.1016/j.aam.2016.05.002
0196-8858/

1. Introduction

Fix a finite alphabet $\mathcal{A}$ and write $|\mathcal{A}|$ for its cardinality. A word is a finite sequence of symbols in the alphabet. A rotation is the operation that moves the final symbol of a word to the first position while shifting all other symbols to the next position, or it is the composition of this operation with itself an arbitrary number of times. A circular word, or necklace, is the equivalence class of a word under rotations. In this note we introduce *perfect necklaces*.

Definition 1. A necklace is (k, n)-*perfect* if it has length $n|\mathcal{A}|^k$ and each word of length k occurs exactly n times at positions which are different modulo n for any convention on the starting point. A necklace is *perfect* if it is (k, k)-perfect for some k.

Perfect necklaces are a variant of the celebrated de Bruijn necklaces [10]. Recall that a de Bruijn necklace of order k in alphabet $\mathcal{A}$ has length $|\mathcal{A}|^k$ and each word of length k occurs in it exactly once. Thus, our $(k, 1)$-perfect necklaces coincide with the de Bruijn necklaces of order k. For a supreme presentation of de Bruijn necklaces, including a historic account of their discovery and rediscovery, see [5]. Observe that a necklace of length $k|\mathcal{A}|^k$ admits k possible decompositions into $|\mathcal{A}|^k$ consecutive (non-overlapping) words of length k. Hence, a necklace is (k, k)-perfect if and only if it has length $k|A|^k$ and each word of length k occurs exactly once in each of the k possible decompositions.

For each k and n, we give a characterization of (k, n)-perfect necklaces in terms of Eulerian circuits in appropriate graphs (Corollary 14). We give a closed formula for the number of (k, n)-perfect necklaces (Theorem 20). These are the most elaborate results in this work.

We show that each arithmetic sequence with difference coprime with the alphabet size induces a perfect necklace (Theorem 5). In particular, the concatenation of all words of the same length in lexicographic order yields a perfect necklace (Corollary 6). This provides a gracious instance of a perfect necklace for any word length.

The combinatorial properties of the concatenation of all words of the same length in lexicographic order were, as far as we know, considered first by É. Barbier [3,2] (see also [1]). Later Champernowne [8] considered them in his construction of a real number normal to base 10, a property defined by Émile Borel [6]. Champernowne worked with alphabet $\mathcal{A} = \{0, 1, \ldots, 9\}$ and for each k, he bounded the number of occurrences of each word of length up to k in the concatenation of all words of length k in lexicographic order. But neither Barbier nor Champernowne mentioned that each word of length k occurs in this sequence exactly k times, once in each of the k different shifts.

2. Perfect necklaces

Notation We write $\mathcal{A}^*$ for the set of all words, and $\mathcal{A}^k$ for the set of all words of length k. The length of a word w is denoted with $|w|$ and the positions in w are numbered from 0

to $|w|-1$. We write $w(i)$ to denote the symbol in the i-th position of w. Let $\theta : \mathcal{A}^* \to \mathcal{A}^*$ be the *shift* operator, such that for each position i, $(\theta w)(i) = w((i+1) \mod |s|)$. That is, the shift operator is defined with the convention of periodicity. We let θ^n denote the application of the shift n times to the right, and with θ^{-n}, n times to the left. As already stated, a necklace is the equivalence class of a word under rotations. To denote a necklace we write $[w]$ where w is any of the words in the equivalence class. For example, if $\mathcal{A} = \{0, 1\}$,

[000] contains a single word 000, because for every n, $\theta^n(000) = 000$.

[110] contains three words $\theta^0(110) = 110$, $\theta^1(110) = 101$ and $\theta^2(110) = 011$.

Example 2. Let $\mathcal{A} = \{0, 1\}$. We add spaces in the examples just for readability.
For words of length 2 there are just two perfect necklaces:

[00 01 10 11],

[00 10 01 11].

This is a perfect necklace for word length 3:

[000 110 101 111 001 010 011 100].

The following are not perfect,

[00 01 11 10],

[000 101 110 111 010 001 011 100].

The so-called *Gray numbers* are not perfect, for instance,

[000 001 011 010 110 111 101 100].

2.1. Each ordered necklace is perfect

Definition 3. For an ordered alphabet $\mathcal{A}$ and a positive integer k, the *k-ordered necklace* has length $k|\mathcal{A}|^k$ and it is obtained by the concatenation of all words of length k in lexicographic order.

For $\mathcal{A} = \{0, 1\}$ the following are the ordered necklaces for k equal to 1, 2 and 3 respectively:

[01],

[00 01 10 11],

[000 001 010 011 100 101 110 111].

We will prove that for every word length, the ordered necklace is perfect. We say that a bijection $\sigma : \mathcal{A}^k \to \mathcal{A}^k$ is a *cycle* if for each $w \in \mathcal{A}^k$ the set $\{\sigma^j(w) : 0 \leq j < |\mathcal{A}|^k\}$ equals $\mathcal{A}^k$. For a word w we write $w(i \dots j)$ to denote the subsequence of w from position i to j.

Lemma 4. *Let $\mathcal{A}$ be a finite alphabet, $\sigma : \mathcal{A}^k \to \mathcal{A}^k$ a cycle and v any word in $\mathcal{A}^k$. Let $s = \sigma^0(v)\sigma^1(v) \dots \sigma^{|\mathcal{A}|^k-1}(v)$. The necklace $[s]$ is perfect if and only if for every ℓ such that $0 \leq \ell < k$, for every $x \in \mathcal{A}^\ell$ and every $y \in \mathcal{A}^{k-\ell}$, there is a unique $w \in \mathcal{A}^k$ such that $w(k-\ell \dots k-1) = x$ and $(\sigma(w))(0 \dots k-\ell-1) = y$.*

Proof. Assume $[s]$ is (k,k)-perfect. Take ℓ such that $0 \leq \ell < k$, $x \in \mathcal{A}^\ell$ and $y \in \mathcal{A}^{k-\ell}$. Consider $\theta^{-\ell}s$, the $(-\ell)$-th shift of s. Since $[s]$ is (k,k)-perfect, xy occurs exactly once in the decomposition of $\theta^{-\ell}s$ in consecutive words of length k. Thus, there is a unique word w in the decomposition of s in consecutive words of length k whose last ℓ symbols are equal to x and whose first $k-\ell$ symbols are equal to y. Conversely, suppose $[s]$ is not (k,k)-perfect. Then, there is some ℓ, $0 \leq \ell < k$, such that the decomposition of $\theta^{-\ell}(s)$ contains two equal words of length k. This contradicts that for every $x \in \mathcal{A}^\ell$ and every $y \in \mathcal{A}^{k-\ell}$, there is a unique $w \in \mathcal{A}^k$ such that $w(k-\ell \dots k-1) = x$ and $(\sigma(w))(0 \dots k-\ell-1) = y$. □

Theorem 5. *Consider the alphabet $\mathcal{A} = \{0, \dots, b-1\}$ where b is an integer greater than or equal to 2, a word length k and a positive integer r coprime with b. Identify the elements of $\mathcal{A}^k$ with the set of integers modulo b^k according to representation in base b. Define the word of length kb^k by the juxtaposition of the elements of $\mathcal{A}^k$ corresponding to the arithmetic sequence $0, r, 2r, \dots, (b^k-1)r$. Then the associated necklace is perfect.*

Proof. Since r is coprime with b, the addition of r defines a cycle $\sigma : \mathcal{A}^k \to \mathcal{A}^k$. We must check that it satisfies the condition in Lemma 4. For any w such that $w(k-\ell \dots k-1) = x$ we have $\sigma(w)(k-\ell \dots k-1) = \tilde{x}$, where abusing notation $\tilde{x} = x + r \mod b^\ell$. Since the word $y\tilde{x}$ appears only one time in the cycle, this fixes a unique $w = \sigma^{-1}(y\tilde{x})$ with $w(k-\ell \dots k-1) = x$ and $(\sigma(w))(0 \dots k-\ell-1) = y$. □

Corollary 6. *For an ordered alphabet $\mathcal{A}$ and word length k, the k-ordered necklace is perfect.*

Proof. Take $r = 1$ in Theorem 5. □

The following proposition is immediate, so we state it without proof.

Proposition 7. *The following operators $\phi : \mathcal{A}^* \to \mathcal{A}^*$ are well defined on necklaces and preserve perfection. That is, for every k and n and for every $s \in \mathcal{A}^*$, if $[s]$ is (k,n)-perfect then $[\phi s]$ is (k,n)-perfect.*

1. *The digit permutation operator defined by $\phi(x_0 \dots x_{kb^k-1}) = (\pi x_0 \dots \pi x_{kb^k-1})$ for any permutation $\pi : \mathcal{A} \to \mathcal{A}$.*
2. *The reflection operator $\phi(x_0 \dots x_{kb^k-1}) = (x_{kb^k-1} \dots x_0)$.*

3. Characterizing and counting perfect necklaces

To characterize and count (k,n)-perfect necklaces in alphabet $\mathcal{A}$ we consider Eulerian circuits in an appropriate directed graph, defined from $\mathcal{A}$, k and n. Recall that an Eulerian circuit in a graph is a path that uses all edges exactly once. A thorough presentation of

the material on graphs that we use in this section can be read in the monographs [12, 19,9]. For the material on combinatorics on words see the books [16,17].

We write $m|n$ when m divides n and we write $\gcd(m,n)$ for the maximum common divisor between m and n.

Definition 8. Let $\mathcal{A}$ be an alphabet with cardinality b, let s be a word length and let n be a positive integer. We define the *astute graph* $G_{s,n}$ as the directed graph, with nb^s nodes, each node is a pair (u,v), where u is in $\mathcal{A}^s$ and v is a number between 0 and $n-1$. There is an edge from (u,v) to (u',v') if the last $s-1$ symbols from u coincide with the first $s-1$ symbols from u' and $(v+1) \mod n = v'$. Observe that $G_{s,n}$ is strongly regular (all nodes have in-degree and out-degree equal to b) and it is strongly connected (there is a path from every node to every other node).

Remark 9. For any alphabet size, the astute graph $G_{k-1,1}$ coincides with a de Bruijn graph of words of length $k-1$; hence, the Eulerian circuits in $G_{k-1,1}$ yield exactly the de Bruijn necklaces of order k.

Although each Eulerian circuit in the astute graph $G_{k-1,n}$ gives one (k,n)-perfect necklace, each (k,n)-perfect necklace can come from several Eulerian circuits in this graph.

3.1. From perfect necklaces to Eulerian circuits

Hereafter, we fix an alphabet $\mathcal{A}$ and we write b for its cardinality.

Definition 10. For a necklace of length ℓ, $[a_0, a_2, \dots a_{\ell-1}]$, we define its *period* as the minimum integer L such that for every non-negative integer j, $a_{j \mod \ell} = a_{(j+L) \mod \ell}$. Notice that the period L always exists, and necessarily $L|\ell$. If the period coincides with the length we say the necklace is *irreducible*.

Definition 11. Let m,n be positive integers. We define $d_{m,n} = \prod p_i^{\alpha_i}$ where $\{p_i\}$ is the set of primes that divide m, and α_i is the exponent of p_i in the factorization of n.

Proposition 12. *The period L of a (k,n)-perfect necklace satisfies the following:*

1. $L = jb^k$ *for* $j|n$.
2. $d_{b,n}|j$.
3. *The corresponding irreducible necklace of length $L = jb^k$ is (k,j)-perfect.*

Proof. Let $[s]$ be (k,n)-perfect, with $s = a_0 \dots a_{nb^k-1}$.

1. Since $[s]$ has length nb^k, we know $L|nb^k$. Let us verify that $b^k|L$. Since $[s]$ has period L, $[a_0 \dots a_{L-1}]$ is a necklace where all words of length k occur the same number of times. Otherwise, it would be impossible that they occur the same number of times

in $[s]$. If each word of length k occurs j times in $[a_0 \dots a_{L-1}]$, then $L = jb^k$. Since $jb^k | nb^k$, we conclude $j|n$.

2. The word $a_0 \dots a_{k-1}$ occurs at position 0 in s but also at positions $L, 2L, \dots, (n/j-1)L$. These positions are of the form qjb^k where $0 \le q < n/j$. These numbers must have pairwise different congruences modulo n. Equivalently, the n/j numbers of the form rb^k, where $0 \le q < n/j$, are all pairwise different modulo n. This last condition holds exactly when $\gcd(b^k, n/j) = 1$, which in turn is equivalent to $\gcd(b, n/j) = 1$, which is equivalent to $d_{b,n}|j$.

3. As argued in Point 1, in the necklace $[a_0 \dots a_{L-1}]$ every word of length k occurs the same number of times. If the positions of two occurrences of a given word were equal modulo j then they would be equal modulo n, but this is impossible because $[s]$ is (k,n)-perfect. □

Proposition 13. *Let N be a (k,j)-perfect necklace. If n is such that $d_{b,n}|j|n$ then the necklace of length nb^k obtained by repeating N exactly n/j times is (k,n)-perfect.*

Proof. Let $\tilde{N}$ be obtained by repeating N exactly n/j times. Then each word of length k occurs in $\tilde{N}$ exactly $j \times n/j = n$ times. Take a word w of length k and let $q_1, \dots, q_j$ be integers, each between 0 and $jb^k - 1$, be the positions of the occurrences of w in N for some convention on the starting point. Then, w occurs in $\tilde{N}$ at positions $q_i + jb^k t$, where $0 \le t < n/j$. Assume $q_{i_1} + jb^k t_1 \equiv q_{i_2} + jb^k t_2 \pmod{n}$. Taking modulo j we conclude $i_1 = i_2$ because N is (k,j)-perfect. Then we have $b^k t_1 \equiv b^k t_2 \pmod{n/j}$. Since $d_{b,n}|j$ we have $\gcd(b, n/j) = 1$, so $t_1 \equiv t_2 \pmod{n/j}$, which implies $t_1 = t_2$. □

Corollary 14. *Fix an alphabet of b symbols, with $b \ge 2$. Let k and n be positive integers. An Eulerian circuit in the astute graph $G_{k-1,n}$ induces a (k,n)-perfect necklace. Each (k,n)-perfect necklace of period jb^k corresponds to j different Eulerian circuits in $G_{k-1,j}$. Therefore, the number of Eulerian circuits in the astute graph $G_{k-1,n}$ is*

$$e(n) = \sum_{d_{b,n}|j|n} j\ p(j),$$

where $p(j)$ is the number of irreducible (k,j)-perfect necklaces.

3.2. The number of Eulerian circuits in astute graphs

Let G be a directed graph with n nodes. The adjacency matrix of a graph G is the matrix $A(G) = (a_{i,j})_{i,j=1}^{n}$ where $a_{i,j}$ is the number of edges between node i and node j. The characteristic polynomial [9] of a graph G is defined as

$$\mathcal{P}(G;x) = \text{determinant}(xI - A(G)),$$

where I is the identity matrix of dimension $n \times n$.

The BEST theorem (for the authors Bruijn, van Aardenne-Ehrenfest, Smith and Tutte) gives a product formula for the number of Eulerian circuits in directed graphs.

Lemma 15 *(BEST Theorem [12]). Let G be a regular connected graph with n nodes. Let v be a node of G and let $r(G)$ be the number of spanning trees oriented towards v. The number of Eulerian circuits in G is*

$$r(G) \cdot \prod_{v=1}^{n} (degree(v) - 1)!$$

Lemma 16 *(Hutschenreurther, Proposition 1.4 [9]). Let G be a regular multigraph with n nodes and degree b. For any of its nodes, the number of spanning trees $r(G)$ oriented to it is*

$$r(G) = \frac{1}{n} \frac{\partial}{\partial x} \mathcal{P}(G; x)|_{x=b},$$

where $\frac{\partial}{\partial x}$ is the derivative with respect to x.

Given a graph G, its line-graph $\Gamma(G)$ is a graph such that each node of $\Gamma(G)$ represents an edge of G; and two nodes of $\Gamma(G)$ are adjacent if and only if their corresponding edges share a common node in G.

Lemma 17 *([9]). For any directed graph G, regular and connected,*

$$\mathcal{P}(\Gamma(G); x) = x^{m-n} \mathcal{P}(G; x),$$

where $\Gamma(G)$ is the line-graph of G, m is the number of edges of G and n is the number of nodes of G.

In the next lemma we write λ for the empty word, namely the unique word in $\mathcal{A}^0$.

Lemma 18. *Let b be any alphabet size, k be a word length, and j be an integer such that $gcd(b, k)|j|k$. Let $G_{0,j}$ be the graph with the set of nodes $\{(\lambda, 0), (\lambda, 1), \ldots (\lambda, j-1)\}$, with b edges from (λ, i) to $(\lambda, i + 1 \mod j)$. Then, $\mathcal{P}(G_{0,j}; x) = x^j - b^j$.*

Proof. It is easy to check that $\mathcal{P}(G_{0,j}; x) = det(xI - A(G_{0,j}))$, which is equal to $x^j - b^j$. □

Lemma 19. *Suppose we have an alphabet of b symbols with $b \geq 2$. Let k be a word length and j be a positive integer such that $\gcd(b, k)|j|k$. The number of Eulerian circuits in the astute graph $G_{k-1,j}$ is $(b!)^{jb^{k-1}} b^{-k}$.*

Proof. We write $\Gamma(G)$ to denote the line graph of G. Notice that for every positive s and for every j, $G_{s,j} = \Gamma(G_{s-1,j})$. In this proof the value j will remain fixed.

Since $G_{k-1,j}$ has jb^{k-1} nodes, each with in-degree b (also out-degree b), by Lemma 15 the number of Eulerian circuits in $G_{k-1,j}$ is

$$r(G_{k-1,j}) \cdot \prod_{v=1}^{jb^{k-1}} (degree(v) - 1)! = r(G_{k-1,j}) \cdot (b-1)!^{jb^{k-1}}.$$

The rest of the proof is devoted to determine $r(G_{k-1,j})$ using Lemma 16.

$$\begin{aligned}
\mathcal{P}(G_{k-1,j}; x) &= \mathcal{P}(\Gamma(G_{k-2,j}); x) \\
&= x^{b^{k-1}j - b^{k-2}j} \mathcal{P}(G_{k-2,j}; x) \\
&= x^{j(b^{k-1} - b^{k-2})} \mathcal{P}(\Gamma(G_{k-3,j}); x) \\
&= x^{j(b^{k-1} - b^{k-2})} x^{j(b^{k-2} - b^{k-3})} \mathcal{P}(G_{k-3,j}; x) \\
&= x^{j(b^{k-1} - b^{k-3})} \mathcal{P}(G_{k-3,j}; x) \\
&= \dots \\
&= x^{j(b^{k-1} - b^0)} \mathcal{P}(G_{0,j}; x) \\
&= x^{j(b^{k-1} - 1)} (x^j - b^j). \\
\frac{\partial}{\partial x} \mathcal{P}(G_{k-1,j}; x) &= \frac{\partial}{\partial x} x^{j(b^{k-1} - 1)} (x^j - b^j) \\
&= (jb^{k-1} - j) x^{jb^{k-1} - j - 1} (x^j - b^j) + x^{jb^{k-1} - j} j x^{j-1}. \\
\frac{\partial}{\partial x} \mathcal{P}(G_{k-1,j}; x)|_{x=b} &= b^{jb^{k-1} - j} j b^{j-1}.
\end{aligned}$$

Finally, by Lemma 16,

$$r(G_{k-1,j}) = \frac{1}{jb^{k-1}} \frac{\partial}{\partial x} \mathcal{P}(G_{k-1,j}; x)|_{x=b} = \frac{1}{jb^{k-1}} b^{jb^{k-1} - j} j b^{j-1} = b^{jb^{k-1} - k}.$$

Hence, the total number Eulerian circuits in $G_{k-1,j}$ is

$$b^{jb^{k-1} - k} ((b-1)!)^{jb^{k-1}} = b!^{jb^{k-1}} b^{-k}. \qquad \square$$

3.3. The number of perfect necklaces

Recall that by Definition 11, $d_{b,n} = \prod p_i^{\alpha_i}$, where $\{p_i\}$ is the set of primes that divide both b and n, and α_i is the exponent of p_i in the factorization of n. The Euler totient function $\varphi(n)$ counts the positive integers less than or equal to n that are relatively prime to n.

Theorem 20. *Suppose we have an alphabet of b symbols, with $b \geq 2$. Let k and n be positive integers. The number of (k,n)-perfect necklaces is*

$$\frac{1}{n} \sum_{d_{b,n}|j|n} e(j)\varphi(n/j)$$

where $e(j) = (b!)^{jb^{k-1}} b^{-k}$ is the number of Eulerian circuits in graph $G_{k-1,j}$ and φ is Euler's totient function.

Proof. Let $p(j)$ be the number of irreducible (k,j)-perfect necklaces. Then, the number of (k,n)-perfect necklaces is

$$\sum_{d_{b,n}|j|n} p(j).$$

Let $e(j)$ be the number of Eulerian circuits in the astute graph $G_{k-1,j}$. By Corollary 14, for each j such that $d_{b,n}|j|n$,

$$e(j) = \sum_{d_{b,n}|\ell|j} \ell \; p(\ell).$$

Notice that $d_{b,n} = d_{b,j}$. For a lighter notation, in the rest of the proof we abbreviate $d_{b,n}$ as just d. Then, writing each such j as a multiple of d, we obtain that for each m such that $md|n$,

$$e(md) = \sum_{i|m} id \; p(id).$$

Let $g(m) = e(md)$ and $f(m) = p(md)\; md$. Writing μ for the Möbius function we obtain

$$\begin{aligned}
f(m) &= \sum_{i|m} \mu(m/i)\; g(i).\\
p(md)\; md &= \sum_{i|m} \mu(m/i)\; e(id).\\
p(md) &= \frac{1}{md} \sum_{i|m} \mu(m/i)\; e(id).\\
\sum_{d|j|n} p(j) &= \sum_{m|n/d} \frac{1}{md} \sum_{i|m} \mu(m/i)\; e(id)\\
&= \sum_{i|n/d} e(id) \sum_{i|m|n/d} \frac{1}{md}\; \mu(m/i)\\
&= \sum_{d|j|n} e(j) \sum_{j|q|n} \frac{1}{q}\; \mu(q/j).
\end{aligned}$$

Applying the Möbius inversion,

$$\sum_{j|q|n} \frac{1}{q}\ \mu(q/j) = \sum_{r|n/j} \frac{1}{jr}\ \mu(r) = \frac{1}{n} \sum_{r|n/j} \frac{n/j}{r}\ \mu(r) = \frac{1}{n}\ \varphi(n/j).$$

We have used the identity $\varphi(m) = \sum_{r|m} \frac{m}{r}\mu(r)$, which is simply the inversion of $m = \sum_{r|m} \varphi(r)$. By Lemma 19, the number $e(j)$ of Eulerian circuits in the astute graph $G_{k-1,j}$ is $(b!)^{jb^{k-1}} b^{-k}$. □

4. Finite-size tests and perfect necklaces

"Given a finite family of tests for randomness there is an infinite sequence x which passes all of them, but x will be rejected by a new more refined test", proposed Norberto Fava to us. Our attempt to formalize this claim led to finite-size tests and perfect periodic sequences. The result is summarized in Proposition 21.

Let $(X_0, X_1, \dots)$ be a sequence of random variables with values in a given alphabet $\mathcal{A}$ with at least two symbols. We say that the sequence is *random* if the variables are uniformly distributed in $\mathcal{A}$ and mutually independent. To test if a sample $(x_0, \dots, x_{n-1}) \in \mathcal{A}^n$ comes from a random sequence we consider the following finite-size hypothesis testing setup. As usual, we write $\mathbb{R}$ for the set of real numbers.

(a) The *hypothesis*

$$H_0 : \ \ (X_0, X_1, \dots) \text{ is random}$$

(b) A *test-size* k and a *test function* $t : \mathcal{A}^k \to \mathbb{R}$. Denote

$$\tau = E_0\big[t(X_0, \dots, X_{k-1})\big] = |\mathcal{A}|^{-k} \sum_{(y_0,\dots,y_{k-1})\in\mathcal{A}^k} t(y_0, \dots, y_{k-1}),$$

where E_0 is the expectation associated with the hypothesis H_0.

(c) A function $T_n : \mathcal{A}^n \to \mathbb{R}$ defined by

$$T_n(x_0, \dots, x_{n-1}) = \Big|\frac{1}{n} \sum_{i=0}^{n-1} t(x_i, \dots, x_{i+k-1}) - \tau\Big|$$

with periodic boundary conditions $x_{n+j} = x_j$. Thus, $T_n(x_0, \dots, x_{n-1})$ is the absolute difference between the empirical mean of t for the sample and the expected value of t under H_0.

(d) An *error* $\varepsilon > 0$ and the *decision rule*

If $T_n(x_0, \dots, x_{n-1}) > \varepsilon$ then reject the sample $(x_0, \dots, x_{n-1})$ as coming from H_0.

In this case we say that *the test t rejects the sample* $(x_0, \dots, x_{n-1})$.

This is called a test of size k because rejection is decided as a function of the empirical mean of t, a function of k successive coordinates. Examples of finite-size tests include frequency test, block testing, number of runs in a block, longest run of ones in a block, etc. There are many (non-finite) tests, like the discrete Fourier transform test, the Kolmogorov–Smirnov test and many others. These tests also use some function $\tilde{T}_n$ of the sample, not necessarily based on the empirical mean of a t. The common feature is the use of the distribution of $\tilde{T}_n(X_1, \dots, X_n)$ under H_0 to compute the probability of rejection when H_0 holds.

Tests for H_0 are used to check if a sequence of numbers produced by a random number generator can be considered random; see Knuth [13] and the battery of tests proposed by L'Ecuyer and Simard [14]. A nice account of the history of hypothesis testing is given by Lehmann [15].

In the usual hypotheses testing the sample-size n is kept fixed. Assuming H_0 and repeating the test j times with independent data, the proportion of times that the hypothesis is rejected converges as $j \to \infty$ to the probability under H_0 that $T_n(X_0, \dots, X_{n-1}) > \varepsilon$. Instead, we will take one infinite sequence, test its first n elements, record rejection for each n and let $n \to \infty$.

Let $x = (x_0, x_1, \dots)$ be an infinite sequence of symbols in $\mathcal{A}$. Fix a test-size k, a test-function t of size k and let T_n be given by (c). We say that x *passes the test* t if

$$\lim_{n\to\infty} T_n(x_0, \dots, x_{n-1}) = 0. \tag{$*$}$$

That is, for each $\varepsilon > 0$ there is an $n(x, \varepsilon)$ such that for all $n > n(x, \varepsilon)$ we have

$$T_n(x_0, \dots, x_{n-1}) \leq \varepsilon.$$

In other words, fixing the test function t of size k and the error ε, the test t rejects $(x_0, \dots, x_{n-1})$ for at most a finite number of n's. When $(*)$ does not hold we say that t *rejects* x.

The random sequence $(X_0, X_1, \dots)$ of independently identically distributed uniform random variables in $\mathcal{A}$ passes any finite-size test t almost surely. This is the same as saying that the set of real numbers in $[0, 1]$ whose $|\mathcal{A}|$-ary representation passes all finite tests has Lebesgue measure 1.

We say that the infinite sequence x is (k, m)-*perfect* if x is periodic with period $m|\mathcal{A}|^k$ and the necklace $[x_0 \dots x_{m|\mathcal{A}|^k-1}]$ is (k, m)-perfect. Recall that $(k, 1)$-perfect necklaces are exactly the de Bruijn necklaces of order k, so the following proposition considers infinite de Bruijn sequences of order k as a special case: if x is de Bruijn of order k there is a test of size $k + 1$ that rejects x.

Proposition 21. *Assume alphabet $\mathcal{A}$ has at least two symbols. Let m be a positive integer and let the infinite sequence x be (k, m)-perfect. Then, the following holds:*

1. *The infinite sequence x passes every test of size $j \leq k$.*
2. *For each $h > k + \log_{|A|} m$ there exists a test t of size h such that t rejects x.*

Proof. Let b be the number of symbols in $\mathcal{A}$. Thus, the period of x has length mb^k.

1. Let t be a test of size k. For any positive integer ℓ, by periodicity,

$$T_{mb^k\ell} = \left|\frac{1}{mb^k\ell}\sum_{i=0}^{mb^k\ell-1} t(x_i,\dots,x_{i+k-1}) - \tau\right| = \left|\frac{\ell}{mb^k\ell}\sum_{i=0}^{mb^k-1} t(x_i,\dots,x_{i+k-1}) - \tau\right| = 0$$

because x is (k,m)-perfect and the definition of τ in (b). Now take $j \in \{0,\dots,mb^k-1\}$ and use the above identity to get

$$(mb^k\ell + j)T_{mb^k\ell+j} = j\,T_j \leq j\max|t-\tau| \leq mb^k \max|t-\tau|,$$

where $\max|t-\tau| = \max_{z_0,\dots,z_{k-1}} |t(z_0,\dots,z_{k-1}) - \tau|$. Hence,

$$T_{mb^k\ell+j} \leq \frac{mb^k}{mb^k\ell + j}\max|t-\tau| \leq \frac{1}{\ell}\max|t-\tau| \underset{\ell\to\infty}{\longrightarrow} 0.$$

This shows that x passes t. Let $\tilde{t}$ be a test of size $j < k$. To see that x also passes $\tilde{t}$ define t of size k as

$$t(x_0,\dots,x_{k-1}) = \tilde{t}(x_0,\dots,x_{j-1}).$$

2. Let h be an integer such that $h > k + \log_b m$. Then $b^h > mb^k$ and there are more words $w = w_0\dots w_{h-1} \in \mathcal{A}^h$ than the possible mb^k places to start. Hence, there is at least one word $\tilde{w}$ of length h not present in the sequence x and the test t consisting on the indicator of $\tilde{w}$ rejects x. □

Finite tests and normal numbers As stated by Borel (see [7]), a real number is simply normal to base b^k exactly when each block of length k occurs in the b-ary expansion of x with asymptotic frequency b^{-k}. Hence, a real number is simply normal to base b^k if its b-ary expansion passes all tests up to size k. We have obtained that for each k and b, and for any m, each (k,m)-perfect sequence in alphabet $\{0,1,\dots,b-1\}$ is the b-ary expansion of a number that is simply normal to base b^k. Borel defines normality to base b as simple normality to all bases b^k, for every positive integer k. Henceforth, a number is normal to base b if its b-ary expansion passes all statistical tests of finite size. Then, each instance of a number normal to a given base provides an example of a sequence that passes all finite-size tests. Many are known, such as [8,4] and the references in [7].

Infinite tests and algorithmically random sequences Martin-Löf introduced tests defined in terms of computability [18], which properly include all tests of finite size. The infinite

sequences that pass all those tests are called Martin-Löf random sequences or algorithmically random sequences. Due to the nature of the definition, the algorithmically random sequences can not be computed but some of them can be defined at the first level of the Arithmetical Hierarchy [11]. Since for every k and m, each (k, m)-perfect sequence is rejected by some Martin-Löf test, (k, m)-perfect sequences are not algorithmically random.

Acknowledgments

We thank Norberto Fava and Victor Yohai for motivating the question on the existence of periodic sequences that pass any finite family of finite-size tests. We thank Liliana Forzani and Ricardo Fraiman for enlightening discussions. We are grateful to an anonymous referee for the reference to the early work of Ém. Barbier.

Alvarez and Becher are members of Laboratoire International Associé INFINIS, Université Paris Diderot–CNRS/Universidad de Buenos Aires–CONICET. Alvarez is supported by CONICET doctoral fellowship. Becher and Ferrari are supported by the University of Buenos Aires and by CONICET.

References

[1] Jean-Paul Allouche, Jeffrey Shallit, Automatic Sequences: Theory, Applications, Generalizations, Cambridge University Press, Cambridge, 2003.
[2] Ém. Barbier, On suppose écrite la suite naturelle des nombres; quel est le $(10^{10000})^{\text{ième}}$ chiffre écrit?, C. R. Séances Acad. Sci. Paris 105 (1887) 1238–1239.
[3] Ém. Barbier, On suppose écrite la suite naturelle des nombres; quel est le $(10^{1000})^{\text{ième}}$ chiffre écrit?, C. R. Séances Acad. Sci. Paris 105 (1887) 795–798.
[4] Verónica Becher, Pablo Ariel Heiber, On extending de Bruijn sequences, Inform. Process. Lett. 111 (18) (2011) 930–932.
[5] Jean Berstel, Dominique Perrin, The origins of combinatorics on words, European J. Combin. 28 (3) (2007) 996–1022.
[6] Émile Borel, Les probabilités dénombrables et leurs applications arithmétiques, Rend. Circ. Mat. Palermo, Suppl. 27 (1909) 247–271.
[7] Yann Bugeaud, Distribution Modulo One and Diophantine Approximation, Cambridge Tracts in Mathematics, vol. 193, Cambridge University Press, Cambridge, 2012.
[8] David Champernowne, The construction of decimals normal in the scale of ten, J. Lond. Math. Soc. S1-8 (4) (1933) 254.
[9] Dragoš M. Cvetković, Michael Doob, Horst Sachs, Spectra of Graphs: Theory and Application, Pure and Applied Mathematics, vol. 87, Academic Press, Inc., Harcourt Brace Jovanovich, Publishers, New York–London, 1980.
[10] Nicolaas G. de Bruijn, A combinatorial problem, Proc. K. Ned. Akad. Wet. 49 (1946) 758–764, Indag. Math. 8 (1946) 461–467.
[11] Rodney G. Downey, Denis R. Hirschfeldt, Algorithmic Randomness and Complexity. Theory and Applications of Computability, Springer, New York, 2010.
[12] Frank Harary, Graph Theory, Addison-Wesley Publishing Co., Reading, Mass.–Menlo Park, Calif.–London, 1969.
[13] Donald E. Knuth, The Art of Computer Programming. Vol. 2: Seminumerical Algorithms, third edition, Addison-Wesley, 1998.
[14] Pierre L'Ecuyer, Richard Simard, TestU01: a C library for empirical testing of random number generators, ACM Trans. Math. Software 33 (4) (2007) 40, Art. 22.
[15] Erich L. Lehmann, Fisher, Neyman, and the Creation of Classical Statistics, Springer, New York, 2011.

[16] M. Lothaire, Combinatorics on Words, Cambridge Mathematical Library, Cambridge University Press, Cambridge, 1997.
[17] M. Lothaire, Algebraic Combinatorics on Words, Encyclopedia of Mathematics and Its Applications, vol. 90, Cambridge University Press, Cambridge, 2002.
[18] Per Martin-Löf, The definition of random sequences, Inf. Control 9 (1966) 602–619.
[19] William T. Tutte, Graph Theory, Encyclopedia of Mathematics and Its Applications, vol. 21, Addison-Wesley Publishing Company, Advanced Book Program, Reading, MA, 1984.