HIDDEN WORD STATISTICS

PHILIPPE FLAJOLET[(1)], WOJCIECH SZPANKOWSKI[(2)], BRIGITTE VALLÉE[(3,1)]

ABSTRACT. We consider the sequence comparison problem, also known as "*hidden*" pattern problem, where one searches for a given *subsequence* in a text (rather than a string understood as a sequence of consecutive symbols). A characteristic parameter is the number of occurrences of a given pattern w of length m as a subsequence in a random text of length n generated by a memoryless source. Spacings between letters of the pattern may either be constrained or not in order to define valid occurrences. We determine the mean and the variance of the number of occurrences, and establish a Gaussian limit law and large deviations. These results are obtained via combinatorics on words, formal language techniques, and methods of analytic combinatorics based on generating functions. The motivation to study this problem comes from an attempt at finding a reliable threshold for intrusion detections, from textual data processing applications, and from molecular biology.

1. INTRODUCTION

String matching and *sequence comparison* are two basic problems of pattern matching known informally as "stringology". Hereafter, by a string we mean a sequence of consecutive symbols. In string matching, given a pattern $\mathcal{W} = w_1w_2\dots w_m$ (of length m) one searches for some/all occurrences of w as a block of consecutive symbols in a text $T = t_1t_2\dots t_n$ (of length n). The algorithms by Knuth–Morris–Pratt and Boyer–Moore [11] provide efficient ways of finding such occurrences. Accordingly, the number of string occurrences in a random text has been intensively studied over the last two decades, with significant progress in this area being reported [5, 20, 21, 32, 33, 34, 41]. For instance Guibas and Odlyzko [20, 21] have revealed the fundamental rôle played by autocorrelation vectors and their associated polynomials. Régnier and Szpankowski [33, 34] established that the number of occurrences of a string is asymptotically normal under a diversity of models that include Markov chains. Nicodème, Salvy, and Flajolet [32] showed generally that the number of places in a random text at which a 'motif' (i.e., a general regular expression pattern) terminates is asymptotically normally distributed.

In sequence comparisons, we search for a given pattern $\mathcal{W} = w_1w_2\dots w_m$ in the text $T_n = t_1t_2\dots t_n$ as a *subsequence*, that is, we look for indices $1 \le i_1 < i_2 < \cdots < i_m \le n$ such that $t_{i_1} = w_1$, $t_{i_2} = w_2$, $\cdots$, $t_{i_m} = w_m$. We also say that the word w is *"hidden"* in the text; thus we call this the *hidden pattern* problem. For example,

Date: October 10, 2002.

(*) This research was supported in part by sponsors of CERIAS at Purdue under contract 1419991431A, by the ALCOM-FT Project (# IST-1999-14186) of the European Union, and by NSF Grants CCR-9804760 and CCR-0208709.

(1) Algorithms Project, INRIA-Rocquencourt, 78153 Le Chesnay, France.

(2) Dept. Computer Science, Purdue University, W. Lafayette, IN 47907, U.S.A.

(3) GREYC, Université de Caen, F-14032 Caen Cedex, France.

`date` occurs as a subsequence in the text `hidden pattern`, in fact four times, but not even once as a string. We allow the possibility of imposing an additional set of constraints $\mathcal{D}$ on the indices $i_1, i_2, \dots, i_m$ to record a valid subsequence occurrence: for a given family of integers d_j ($d_j \geq 1$, possibly $d_j = \infty$), one should have $(i_{j+1} - i_j) \leq d_j$. In other words, the allowed lengths of the "gaps" $(i_{j+1} - i_j - 1)$ should be $< d_j$. With $\#$ representing a 'don't-care-symbol' (similar to the unix '$\star$'-convention) and the subscript denoting a strict upper bound on the length of the associated gap, a typical pattern may look like

$$\mathtt{ab}\#_2\mathtt{r}\#\mathtt{ac}\#\mathtt{a}\#\mathtt{d}\#_4\mathtt{a}\#\mathtt{br}\#\mathtt{a}; \tag{1}$$

there, $\#$ abbreviates $\#_\infty$ and $\#_1$ (which glues adjacent letters together) is omitted; the meaning is that '`ab`' should occur first contiguously, followed by '`r`' with a gap of < 2 symbols, followed anywhere later in the text by '`ac`', etc. The case when all the d_j's are infinite is called the (fully) *unconstrained problem*. When all the d_j's are finite, we speak of the (fully) *constrained* problem—in particular, the case where all d_j reduce to 1 gives back classical string matching as a limit case.

Motivations. Our original motivation to study this problem came from *intrusion detection* in the area of computer security. The problem is important due to the rise of attacks on computer systems. There are several approaches to intrusion detections, but, recently the pattern matching approach has found many advocates, most notably in [3, 31, 42]. The main idea of this approach is to search in an audit file (the text) for certain patterns (then known as "signatures") representing suspicious activities that might be indicative of an intrusion by an outsider, or misuse of the system by an insider. The key to this approach is to recognize that these patterns are **subsequences** because an intrusion signature specification requires the possibility of a variable number of intervening events between successive events of the signature. In practice one often needs to put some additional restrictions on the distance between the symbols in the searched subsequence, which leads to constrained version of subsequence pattern matching. The fundamental question is then: *How many occurrences of a signature (subsequence) indicate a real attack?* In other words, how does one set a *threshold* so that real intrusions are detected and false alarms are avoided? It is clear that *random* (unpredictable) events occur and setting the threshold too low will lead to an unrealistic number of false alarms. On the other hand, setting the threshold too high may result in missing some attacks, which is perhaps even more dangerous. This fundamental problem initially motivated our studies of hidden pattern statistics. By knowing the most likely number of occurrences and the probability of deviating from it, we can set a threshold such that with a small probability we miss real attacks (cf. Remark after Theorem 3 for a solution).

Molecular biology provides another important source of applications [35, 40, 41]. As a rule, there, one searches for subsequences, not strings. Examples are in abundance: split genes where *exons* are interrupted by *introns*, *starting* and *stopping* signal in genes, *tandem repeats* in DNA, etc. In general, for gene searching, the constrained hidden pattern matching (perhaps with an exotic constraint set) is the right approach for finding meaningful information. The hidden pattern problem can also be viewed as a close relative of the longest common subsequence (LCS) problem, itself of immediate relevance to computational biology, but whose probabilistic aspects are still surrounded by mystery [37].

We, computer scientists and mathematicians, are certainly not the first who invented hidden words and hidden meaning [2]. Rabbi Akiva in the first century A.D. wrote a collection of documents called *Maaseh Merkava* on secret mysticism and meditations. In the eleventh century Spanish Solomon Ibn Gabirol called these secret teachings *Kabbalah*. Kabbalists organized themselves as a secret society dedicated to the study of the ancient wisdom of Torah, looking for mysterious connections and hidden truth, meaning, and words in Kaballah and elsewhere (without computers!). Recent versions of this activity are *knowledge discovery and data mining, bibliographic search, lexicographic research, textual data processing*, or even *web site indexing*. Public domain utilities like `agrep`, `grappe`, `webglimpse` (developed by Manber and Wu [43], Kucherov [30], and others) depend crucially on approximate pattern matching algorithms for subsequence detection. Many interesting algorithms, based on regular expressions and automata, dynamic programming, directed acyclic word graphs, digital tries or suffix trees have been developed; see [8, 12, 30, 43] for a flavour of the diversity of approaches to algorithmic design.

In all of the contexts mentioned above, it is of obvious interest to discern what constitutes a meaningful observation of pattern occurrences from what is merely a statistically unavoidable phenomenon (noise!). This is precisely the problem addressed here. We establish *subsequence statistics*, i.e., precise probabilistic information on the number of occurrences of a given pattern $\mathcal{W}$ as a subsequence in a random text T_n generated by a memoryless source—this, in the most general case covering the constrained and unconstrained versions as well as mixed situations. Surprisingly enough and to the best of our knowledge, there are no results in the literature that address the question at this level of generality. An immediate consequence of our results is the possibility to set *thresholds* at which appearance of a (subsequence) pattern starts being meaningful.

Results. Let $\Omega(T)$ be the *number of occurrences* of a given pattern $\mathcal{W}$ as a subsequence in a text T. By number of occurrences is understood the number of ways the pattern together with its distance constraints can be embedded in the text. We investigate the general case where we allow some of the gaps to be restricted, and others to be unbounded. Then the most important parameter is the quantity b defined as the number of constrained blocks, that is, the number of unbounded gaps (the number of indices j for which $d_j = \infty$) plus 1.

Throughout this article, the text is assumed to be generated by a memoryless source, also known as Bernoulli source, i.e., symbols are drawn independently according to some fixed probability distribution over letters of the alphabet. For a random text T of length n under this model, the *combinatorial* parameter $\Omega(T)$ becomes a *random variable* that is then naturally also denoted as Ω. We prove in Theorem 1 that the number of occurrences Ω in a random text of size n has expectation and variance given by

$$\mathbb{E}_n[\Omega] \sim \frac{n^b}{b!} D\, \pi(\mathcal{W}), \qquad \mathbb{V}_n[\Omega] \sim \sigma^2(\mathcal{W})\, n^{2b-1},$$

with $\mathbb{E}$ and $\mathbb{V}$ denoting the mean and variance operators, while the subscript n indexes the probabilistic model of use. There D is the product of all the finite constraints d_j, $\pi(\mathcal{W})$ is the probability of $\mathcal{W}$, and $\sigma^2(\mathcal{W})$ is a computable constant that depends explicitly (though intricately) on the structure of the pattern $\mathcal{W}$ and the constraints. Then we prove the central limit law by moment methods, that is, we show that all centered moments $(\Omega - \mathbb{E}_n[\Omega])/n^{b-\frac{1}{2}}$ converge to the appropriate moments of the Gaussian distribution (Theorem 2). We stress that, except in the constrained case, the difficulty of the analysis lies in a nonlinear growth of the mean and the variance so that many standard approaches to establishing the central limit law tend to fail.

For the (fully) unconstrained problem, one has $b = m$, and both the mean and the variance admit pleasantly simple closed forms. For the (fully) constrained case, one has $b = 1$, while the mean and the variance become of linear growth. To visualize the dependency of $\sigma^2(\mathcal{W})$ on $\mathcal{W}$, we observe that, when all the d_j equal 1, the problem further reduces to traditional *string matching*, which was extensively studied in the past as witnessed by the (incomplete) list of references: [5, 20, 21, 32, 33, 34, 41]. It is well known that for string matching the variance coefficient σ^2 is a function of the so-called *autocorrelation* of the string. In the general case of hidden pattern matching, the autocorrelation must be replaced by a more complex quantity that depends on the way pairs of constrained occurrences may intersect (cf. Theorem 1 and Section 3.3).

Methodology. The way we approach the probabilistic analysis is through a formal description of situations of interest by means of regular languages. Basically such a description of *contexts* of one, two, or several occurrences gives access to expectation, variance, and higher moments, respectively. A systematic translation into *generating functions* is available by methods of analytic combinatorics deriving from the original Chomsky-Schützenberger theorem. Then, the structure of the implied generating functions at the pole $z = 1$ provides the necessary asymptotic information. In fact, there is an important phenomenon of *asymptotic simplification* where the essentials of combinatorial-probabilistic features are reflected by the singular forms of generating functions. For instance, variance coefficients come out naturally from this approach together with, for each case, a suitable notion of correlation; higher moments are seen to arise from a fundamental asymptotic symmetry of the problem, a fact that eventually carries with it the possibility of estimating moments. From there Gaussian laws eventually result by basic moment convergence theorems. Perhaps the originality of the present approach lies in such a joint use of combinatorial-enumerative techniques and of analytic-probabilistic methods.

An extended abstract of this article has appeared in the proceedings of the ICALP'2001 Colloquium [14].

2. Framework

We fix an alphabet $\mathcal{A} := \{a_1, a_2, \dots, a_r\}$. The set of all possible texts is $\mathcal{A}^\star$ (the set of words over the alphabet $\mathcal{A}$), and a text of length n is an element $T = t_1 t_2 \cdots t_n$ of $\mathcal{A}^n$. A particular matching problem is determined by a pair $(\mathcal{W}, \mathcal{D})$ called a "hidden pattern" specification: the *pattern* $\mathcal{W} = w_1 \cdots w_m$ is a word of length m; the *constraint* $\mathcal{D} = (d_1, \dots, d_{m-1})$ is an element of $(\mathbb{N}^+ \cup \{\infty\})^{m-1}$. The case $\mathcal{D} = (\infty, \dots, \infty)$ models the *unconstrained problem*; at the other end of the spectrum, there lies the case where all d_j are finite, which we name the *constrained problem*.

Positions and occurrences. An m-tuple $I = (i_1, i_2, \dots, i_m)$ $(1 \le i_1 < i_2 < \cdots < i_m)$ satisfies the constraint $\mathcal{D}$ if $i_{j+1} - i_j \le d_j$, in which case it is called a *position*. Let $\mathcal{P}_n(\mathcal{D})$ be the set of all positions subject to the separation constraint $\mathcal{D}$, satisfying furthermore $i_m \le n$. Let also $\mathcal{P}(\mathcal{D}) = \bigcup_n \mathcal{P}_n(\mathcal{D})$. An *occurrence* of pattern $\mathcal{W}$ subject to the constraint $\mathcal{D}$ is a pair (I,T) formed with a position $I=(i_1,i_2,\ldots,i_m)$ of $\mathcal{P}_n(\mathcal{D})$ and a text $T=t_1t_2\cdots t_n$ for which $t_{i_1}=w_1, t_{i_2}=w_2,\ldots,t_{i_m}=w_m$. Thus, what we call an occurrence is a text augmented with the distinguished positions at which the pattern occurs. The number Ω of occurrences of pattern w in text T subject to the constraint $\mathcal{D}$ is then a sum of characteristic variables

(2)
$$\Omega(T)=\sum_{I\in\mathcal{P}_{|T|}(\mathcal{D})} X_I(T), \qquad \text{with} \qquad X_I(T):=[\![w \text{ occurs at position } I \text{ in } T]\!].$$

There, Iverson's bracket convention is used:

$$[\![B]\!]=\begin{cases} 1 & \text{if the property } B \text{ holds,}\\ 0 & \text{otherwise.}\end{cases} \tag{3}$$

Blocks and aggregates. In the general case, the subset $\mathcal{F}$ of indices j for which d_j is finite ($d_j<\infty$) has cardinality $m-b$ with $1\le b\le m$. The two extreme values of b, namely, $b=m$ and $b=1$, thus describe the (fully) unconstrained and the (fully) constrained problem respectively. The subset $\mathcal{U}$ of indices j for which d_j is unbounded ($d_j=\infty$) has cardinality $b-1$. It then separates the pattern $\mathcal{W}$ into b independent subpatterns that are called the *blocks* and are denoted by $\mathcal{W}_1,\mathcal{W}_2,\ldots\mathcal{W}_b$. All the possible d_j "inside" any $\mathcal{W}_r$ are finite and form the subconstraint $\mathcal{D}_r$, so that a general hidden pattern specification $(\mathcal{W},\mathcal{D})$ is equivalently described as a b-tuple of fully constrained hidden patterns $((\mathcal{W}_1,\mathcal{D}_1),(\mathcal{W}_2,\mathcal{D}_2),\ldots,(\mathcal{W}_b,\mathcal{D}_b))$.

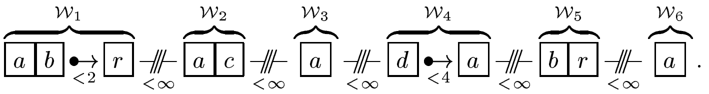
Example. With the example (1) described in the introduction, namely,

$$\texttt{ab}\#_2\texttt{r}\#\texttt{ac}\#\texttt{a}\#\texttt{d}\#_4\texttt{a}\#\texttt{br}\#\texttt{a},$$

one has $b=6$, the six blocks being

$$\mathcal{W}_1=\texttt{a}\#_1\texttt{b}\#_2\texttt{r},\ \mathcal{W}_2=\texttt{a}\#_1\texttt{c},\ \mathcal{W}_3=\texttt{a},\ \mathcal{W}_4=\texttt{d}\#_4\texttt{a},\ \mathcal{W}_5=\texttt{b}\#_1\texttt{r},\ \mathcal{W}_6=\texttt{a}.$$

In more figurative terms, this is described as follows (with springs representing unbounded gaps, representing bounded gaps, and gaps <1 omitted):



The decomposition of the hidden pattern problem into blocks is a fundamental tool in what follows. □

In the same way, an occurrence position $I=(i_1,i_2,\ldots,i_m)$ of $\mathcal{W}$ subject to constraint $\mathcal{D}$ gives rise to b suboccurrences, $I^{[1]},I^{[2]},\ldots I^{[b]}$, the rth term $I^{[r]}$ representing an occurrence of $\mathcal{W}_r$ subject to constraint $\mathcal{D}_r$. The rth *block* $B^{[r]}$ is the closed segment whose end points are the extremal elements of $\mathcal{I}^{[r]}$, and the *aggregate* of position I, denoted by $\alpha(I)$, is the collection of these b blocks. In our example, the position

$$I=(6,7,9,18,19,22,30,33,50,51,60)$$

satisfies the constraint $\mathcal{D}$ and gives rise to six subpositions,

$$\overbrace{(6,7,9)}^{I^{[1]}},\quad \overbrace{(18,19)}^{I^{[2]}},\quad \overbrace{(22)}^{I^{[3]}},\quad \overbrace{(30,33)}^{I^{[4]}},\quad \overbrace{(50,51)}^{I^{[5]}},\quad \overbrace{(60)}^{I^{[6]}}\ ;$$

accordingly, the resulting aggregate $\alpha(I)$,

$$\overbrace{[6,9]}^{B^{[1]}},\quad \overbrace{[18,19]}^{B^{[2]}},\quad \overbrace{[22]}^{B^{[3]}},\quad \overbrace{[30,33]}^{B^{[4]}},\quad \overbrace{[50,51]}^{B^{[5]}},\quad \overbrace{[60]}^{B^{[6]}},$$

is formed with six blocks. □

Finally, for a totally constrained hidden pattern $(\mathcal{W}, \mathcal{D})$, we associate two quantities: the length of a constraint, and the product of a constraint,

$$\ell(\mathcal{D}) = 1 + \sum_i d_i, \qquad D(\mathcal{D}) := \prod_i d_i;$$

this is extended to a general hidden pattern specification as

$$\ell(\mathcal{D}) = \sum_{i=1}^{b} \ell(\mathcal{D}_i), \qquad D(\mathcal{D}) := \prod_{i=1}^{b} D(\mathcal{D}_i).$$

Probabilistic model. As regards the probabilistic model, we consider a *memoryless source* that emits symbols of the text independently from the fixed finite alphabet $\mathcal{A} = \{a_1, a_2, \ldots a_r\}$ and denote by p_α $(0 < p_\alpha < 1)$ the probability of the symbol $\alpha \in \mathcal{A}$ being emitted. For a given length n, a random *text*, denoted by T_n is then drawn according to the Bernoulli model corresponding to the product probability on $\mathcal{A}^n$:

$$\pi(T) \equiv \pi(t_1 \cdots t_n) = \prod_{i=1}^{n} p_{t_i}. \tag{4}$$

The *pattern* $\mathcal{W} = w_1 \cdots w_m$ of length m is fixed, and the quantity $\pi(\mathcal{W}) = \prod_{i=1}^{n} p_{w_i}$, the pattern "probability", surfaces throughout the analysis. Under the randomness model, the restriction of Ω to $\mathcal{A}^n$, denoted by Ω_n whenever dependency on size needs to be made explicit, becomes a *random variable* defined on $\mathcal{A}^n$. Then, Ω_n, is itself a sum of correlated random variables X_I (defined in (2)) for all allowable $I \in \mathcal{P}_n(\mathcal{D})$.

Generally speaking, we shall use in the sequel other parameters defined on $\mathcal{A}^\star$, namely Ξ defined in (11) and $\widetilde{\Xi}$ defined in (24). For any such parameter, say U, we shall adopt similar notations: U is the parameter defined on $\mathcal{A}^\star$, U_n may be used to mark the restriction of U to $\mathcal{A}^n$, and the subscript n appended to $\mathbb{P}, \mathbb{E}, \mathbb{V}$ indicates that the probabilistic model is the product probability on $\mathcal{A}^n$.

Generating functions. We shall consider throughout this paper structures superimposed on words. For a class $\mathcal{C}$ of structures and given a weight function γ from $\mathcal{C}$ to the set of reals, we introduce the *generating function* of the weighted set,

$$C(z) \equiv \sum_n C_n z^n := \sum_{\omega \in \mathcal{C}} \gamma(\omega) z^{|\omega|},$$

where $|\omega|$ denotes the size of structure ω. In particular, the usual counting generating function corresponds to the constant weight $\gamma \equiv 1$. Then[1], $C_n = [z^n]C(z)$ is the total weight of all structures of size n in $\mathcal{C}$.

For structures arising from words the number of letters involved in the structure will determine the size of the structure. The weights will be induced by the probabilities of individual letters. As we shall see in the next section, the collection of occurrences can be described by means of regular expressions extended with disjoint

[1]The notation $[z^n]f(z)$ represents the coefficient of z^n in the series $f(z)$.

unions, and Cartesian products. Thus a minimal set of rules must first be given in order to translate such basic constructions; see [16, 36, 38] for a general framework.

Take $\mathcal{A}, \mathcal{B}, \mathcal{C}$ to be weighted sets with respective weights α, β, γ. Here is a brief summary of translation rules from weighted sets to generating functions:

— *Disjoint unions.* Assume that $\mathcal{C} = \mathcal{A} \cup \mathcal{B}$ where the union is disjoint ($\mathcal{A} \cap \mathcal{B} = \emptyset$), and that the weight γ on $\mathcal{C}$ is inherited from the weights α, β on $\mathcal{A}, \mathcal{B}$:

$$\gamma(\omega) = \begin{cases} \alpha(\omega) & \text{if } \omega \in \mathcal{A} \\ \beta(\omega) & \text{if } \omega \in \mathcal{B}. \end{cases}$$

Then, the corresponding generating functions satisfy

$$C(z) = A(z) + B(z).$$

The proof is a one-liner, given the definitions:

$$C(z) := \sum_{\omega \in \mathcal{C}} \gamma(\omega) z^{|\omega|} = \left(\sum_{\omega \in \mathcal{A}} \alpha(\omega) z^{|\omega|} \right) + \left(\sum_{\omega \in \mathcal{B}} \beta(\omega) z^{|\omega|} \right) =: A(z) + B(z).$$

Disjoint unions in such contexts are also called combinatorial sums and denoted by the symbol '+'.

— *Cartesian products.* Assume that $\mathcal{C} = \mathcal{A} \times \mathcal{B}$ is a Cartesian product and that the weight γ on $\mathcal{C}$ is defined multiplicatively from the weights α, β on $\mathcal{A}, \mathcal{B}$: $\gamma(\langle x, y \rangle) = \alpha(x) \cdot \beta(y)$. Then, the corresponding generating functions satisfy

$$C(z) = A(z) \cdot B(z),$$

since one has

$$C(z) := \sum_{\omega \in \mathcal{C}} \gamma(\omega) z^{|\omega|} = \left(\sum_{\omega \in \mathcal{A}} \alpha(\omega) z^{|\omega|} \right) \cdot \left(\sum_{\omega \in \mathcal{B}} \beta(\omega) z^{|\omega|} \right) =: A(z) \cdot B(z).$$

(A similar translation by products of generating functions holds for unambiguous concatenations of formal languages.)

For an alphabet $\mathcal{A}$ weighted by letter probabilities, the generating function is simply z. By the rules above, the generating function of all words of length n under the Bernoulli model is z^n and the generating function of the entire language $\mathcal{A}^\star$ is

$$1 + z + z^2 + \cdots = \frac{1}{1-z}.$$

As we shall see, the constructions recalled above suffice to express moments of occurrence counts. Consequently, all the resulting generating functions are *rational*, of the special form $F(z) = (1-z)^{-(k+1)} P(z)$ for some integer $k \geq 0$ and polynomial P. This in turn entails precise coefficient asymptotics, namely,

$$[z^n] \frac{P(z)}{(1-z)^{k+1}} = \frac{n^k}{k!} P(1) + O\left(n^{k-1}\right). \tag{5}$$

3. Mean and Variance Analysis

In this section, we assemble definitions and methods described in Section 2 in order to derive estimates of the mean and variance of the number of occurrences (Theorem 1).

3.1. **Mean value analysis.** The first moment of the number of occurrences is easily obtained by describing the collection of all occurrences in terms of formal languages.

We recall that an *occurrence* of pattern $\mathcal{W}$ subject to the constraint $\mathcal{D}$ is a pair (I,T) formed with a position $I=(i_1,i_2,\ldots,i_m)$ of $\mathcal{P}_n(\mathcal{D})$ and a text $T=t_1t_2\cdots t_n$ for which $t_{i_1}=w_1, t_{i_2}=w_2,\ldots,t_{i_m}=w_m$. We consider the collection of position-text pairs

$$\mathcal{O} := \{(I,T) \ ; \ I\in\mathcal{P}_{|T|}(\mathcal{D})\},$$

with the size of an element being by definition the length n of the text T. The weight of an element of $\mathcal{O}$ is taken to be equal to $X_I(T)\pi(T)$. (Here, $\pi(T)$ is the probability of the text), In this way, $\mathcal{O}$ can also be regarded as the collection of all occurrences weighted by probabilities of the text. The corresponding generating function of $\mathcal{O}$ equipped with this weight is

$$O(z) = \sum_{(I,T)\in\mathcal{O}} X_I(T)\pi(T)\, z^{|T|} = \sum_{T}\left(\sum_{I\in\mathcal{P}_{|T|}(\mathcal{D})} X_I(T)\right)\pi(T)z^{|T|}, \tag{6}$$

and, with the definition of Ω,

$$O(z) = \sum_{T}\Omega(T)\pi(T)\, z^{|T|} \ = \sum_{n}\mathbb{E}_n[\Omega]z^n. \tag{7}$$

As a consequence, one has $[z^n]O(z)=\mathbb{E}_n[\Omega]$, so that $O(z)$ serves as the generating function (in the usual sense) of the sequence of expectations $\mathbb{E}_n[\Omega]$.

On the other hand, each occurrence can be viewed as a "context" with an initial string, then the first letter of the pattern, then a separating string, then the second letter, etc. The collection $\mathcal{O}$ is then described combinatorially by

$$\mathcal{O} \ = \mathcal{A}^\star\times\{w_1\}\times\mathcal{A}^{<d_1}\times\{w_2\}\times\mathcal{A}^{<d_2}\times\ldots\times\{w_{m-1}\}\times\mathcal{A}^{<d_{m-1}}\times\{w_m\}\times\mathcal{A}^\star. \tag{8}$$

There, for $d<\infty$, $\mathcal{A}^{<d}$ denotes the collection of all words of length strictly less d, i.e., $\mathcal{A}^{<d} := \bigcup_{i<d}\mathcal{A}^i$, whereas, for $d=\infty$, $\mathcal{A}^{<\infty}$ denotes the collection of all finite words, i.e., $\mathcal{A}^{<\infty} := \mathcal{A}^\star = \bigcup_{i<\infty}\mathcal{A}^i$. Since the source is memoryless, the rules discussed at the end of the last section can be applied, and they give access to $O(z)$ from the description (8). The generating function functions associated to $\mathcal{A}^{<d}$ and $\mathcal{A}^{<\infty}$ are

$$A^{<d}(z) = 1+z+z^2+\cdots+z^{d-1} = \frac{1-z^d}{1-z}, \qquad A^{<\infty}(z) = 1+z+z^2+\cdots = \frac{1}{1-z}.$$

Thus, the description (8) of occurrences automatically translates into

$$O(z) \equiv \sum_{n\geq 0}\mathbb{E}_n[\Omega]\ z^n = \left(\frac{1}{1-z}\right)^{b+1}\times\left(\prod_{i=1}^m p_{w_i}z\right)\times\left(\prod_{i\in\mathcal{F}}\frac{1-z^{d_i}}{1-z}\right). \tag{9}$$

With $\pi(\mathcal{W})$ the probability of the pattern $\mathcal{W}$, one finds finally from (5) and (9):

$$\mathbb{E}_n[\Omega] = [z^n]O(z) = \frac{n^b}{b!}\left(\prod_{i\in\mathcal{F}} d_i\right)\ \pi(\mathcal{W})\left(1+O\left(\frac{1}{n}\right)\right), \tag{10}$$

and a complete asymptotic expansion could be easily obtained. This symbolic derivation of mean values extends the case of standard string matching exposed in [36, p. 366]. Its full significance is revealed when it is applied to higher moment estimates.

3.2. **Variance analysis.** For the analysis of variance and especially of higher moments, it is essential to work with a centered random variable Ξ defined, for each n, as

$$\Xi_n := \Omega_n - \mathbb{E}_n[\Omega] = \sum_{I \in \mathcal{P}_n(\mathcal{D})} Y_I, \qquad \text{with} \qquad Y_I := X_I - \mathbb{E}[X_I] = X_I - \pi(\mathcal{W}). \tag{11}$$

The second moment of the centered variable Ξ equals the variance of Ω and with the centered variables defined above by (11), one has

$$\mathbb{E}_n[\Xi^2] = \sum_{I,J \in \mathcal{P}_n(\mathcal{D})} \mathbb{E}[Y_I Y_J]. \tag{12}$$

From this last equation, we need to analyse *pairs of positions* $(I,T),(J,T) \cong (I,J,T)$ relative to a common text T. We denote by $\mathcal{O}_2$ this set,

$$\mathcal{O}_2 := \{(I,J,T) \ ; \ I,J \in \mathcal{P}_{|T|}(\mathcal{D})\},$$

and we weight each element (I,J,T) by $Y_I(T)Y_J(T)\pi(T)$. The corresponding generating function, which enumerates pairs of occurrences, is

$$O_2(z) := \sum_{(I,J,T) \in \mathcal{O}_2} Y_I(T)Y_J(T)\pi(T)\, z^{|T|} = \sum_T \left(\sum_{I,J \in \mathcal{P}_{|T|}(\mathcal{D})} Y_I(T)Y_J(T) \right) \pi(T) z^{|T|}$$

and, with Equation (12),

$$O_2(z) = \sum_{n \geq 0} \sum_{I,J \in \mathcal{P}_n(\mathcal{D})} \mathbb{E}[Y_I Y_J]\, z^n = \sum_{n \geq 0} \mathbb{E}_n[\Xi^2]\, z^n.$$

The process entirely parallels the derivation of (6) and (7), and, one has $[z^n]O_2(z) = \mathbb{E}_n[\Xi^2]$, so that $O_2(z)$ serves as the generating function (in the usual sense) of the sequence of moments $\mathbb{E}_n[\Xi^2]$.

There are two kinds of pairs (I,J) according as they intersect or not. When I and J do not intersect, the corresponding random variables Y_I and Y_J are independent, and the corresponding covariance $E[Y_I Y_J]$ reduces to 0. As a consequence, one may restrict attention to pairs of occurrences I,J that intersect at one place at least. Suppose that there exist two occurrences of pattern $\mathcal{W}$ at positions I and J which intersect at ℓ distinct places. We then denote by $\mathcal{W}_{I\cap J}$ the subpattern of $\mathcal{W}$ that occurs at position $I \cap J$, and by $\pi(\mathcal{W}_{I\cap J})$ the probability of this subpattern. Since the expectation $\mathbb{E}[X_I X_J]$ equals $\pi(\mathcal{W})^2/\pi(\mathcal{W}_{I\cap J})$, the expectation $\mathbb{E}[Y_I Y_J] = \mathbb{E}[X_I X_J] - \pi(\mathcal{W})^2$ involves a correlation number $e(I,J)$

$$\mathbb{E}[Y_I Y_J] = \pi^2(\mathcal{W})\; e(I,J), \qquad \text{with} \qquad e(I,J) = \frac{1}{\pi(\mathcal{W}_{I\cap J})} - 1. \tag{13}$$

Remark that this relation remains true even if the pair (I,J) is not intersecting, since, in this case, one has $\pi(\mathcal{W}_{I\cap J}) = \pi(\varepsilon) = 1$.

Aggregates and degrees of freedom of pairs of positions. As it turns out in the analysis, asymptotic behaviour is driven by the overlapping of blocks involved in I and J, rather than plainly by the cardinality of $I \cap J$. In order to formalize this, define first the (joint) *aggregate* $\alpha(I,J)$ to be the system of blocks obtained by merging together all intersecting blocks of the two aggregates $\alpha(I)$ and $\alpha(J)$. The number of blocks $\beta(I,J)$ of $\alpha(I,J)$ plays a fundamental rôle here, since it measures the *degree of freedom* of pairs; we also call $\beta(I,J)$ the degree of pair (I,J). Figure 1 illustrates graphically this notion.

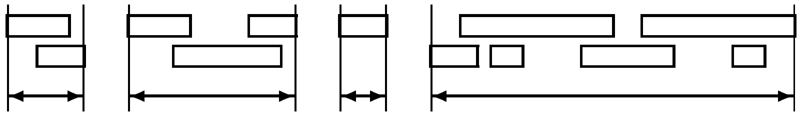


FIGURE 1. A pair of occurrences I, J with $b = 6$ blocks each and the joint aggregates; the number of degrees of freedom is here $\beta(I, J) = 4$.

EXAMPLE. Consider the pattern $\mathcal{W} = \boxed{\mathtt{a}\#_3\mathtt{b}\#_4\mathtt{r}}\,\mathtt{\#}\,\boxed{\mathtt{a}\#_4\mathtt{c}}$ composed of two blocks. The text aarbarbccaracc contains several valid occurrences of $\mathcal{W}$ including two at positions $I = (2, 4, 6, 10, 13)$ and $J = (5, 7, 11, 12, 13)$. The individual aggregates are $\alpha(I) = \{[2, 6], [10, 13]\}, \alpha(J) = \{[5, 11], [12, 13]\}$ so that the joint quantities are: $\alpha(I, J) = [2, 13]$ and $\beta(I, J) = 1$. This pair has exactly degree 1. □

When I and J intersect, there exists at least one block of $\alpha(I)$ that intersects a block of $\alpha(J)$, so that the degree $\beta(I, J)$ is at most equal to $2b - 1$. Next, we partition $\mathcal{O}_2$ according to the value of $\beta(I, J)$ and write

$$\mathcal{O}_2^{[p]} := \{(I, J, T) \in \mathcal{O}_2 \quad ; \quad \beta(I, J) = 2b - p\}$$

for the collection of intersecting pairs (I, J, T) of occurrences for which the degree of freedom equals $2b - p$. From the preceding discussion, only $p \geq 1$ needs to be considered and

$$O_2(z) = O_2^{[1]}(z) + O_2^{[2]}(z) + O_2^{[3]}(z) + \cdots .$$

As we see next, it is only the first term of this sum that matters asymptotically.

Full pairs. In order to conclude the discussion, we need the notion of *full pairs*: a pair (I, J) of $\mathcal{P}_q(\mathcal{D}) \times \mathcal{P}_q(\mathcal{D})$ is *full* if the joint aggregate $\alpha(I, J)$ completely covers the interval $[1, q]$; see Figure 2. (Clearly, the possible values of length q are finite, since q is at most equal to 2ℓ, where ℓ is the length of the constraint $\mathcal{D}$.)

EXAMPLE. Consider the pattern $\mathcal{W} = \mathtt{a}\#_3\mathtt{b}\#_4\mathtt{r}\#\mathtt{a}\#_4\mathtt{c}$. The text aarbarbccaracc also contains two other occurrences of $\mathcal{W}$, at positions $I' = (1, 4, 6, 12, 13)$ and $J' = (5, 7, 11, 12, 14)$. Now, I' and J' are intersecting, and the aggregates are $\alpha(I') = \{[1, 6], [12, 13]\}, \alpha(J') = \{[5, 11], [12, 14]\}$ so that $\alpha(I', J') = \{[1, 11], [12, 14]$. We have here an example of a full pair of occurrences with a number of blocks $\beta(I', J') = 2$. □

There is a fundamental translation invariance due to the independence of symbols in the Bernoulli model that entails a combinatorial isomorphism ($\cong$ represents combinatorial isomorphism)

$$\mathcal{O}_2^{[p]} \cong (\mathcal{A}^\star)^{2b-p+1} \times \mathcal{B}_2^{[p]},$$

where $\mathcal{B}_2^{[p]}$ is the subset of $\mathcal{O}_2$ formed of full pairs such that $\beta(I, J)$ equals $2b - p$. In essence, the gaps can be all grouped together (their number is $2b - p + 1$, which

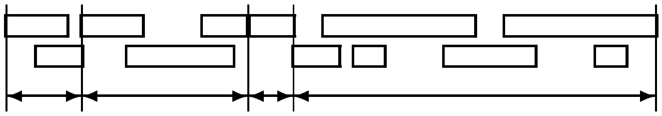


FIGURE 2. A full pair of occurrences I, J with $b = 6$ blocks each.

is translated by the prefactor $(\mathcal{A}^{\star})^{2b-p+1}$), while what remains constitutes a full occurrence. The generating function of $\mathcal{O}_2^{[p]}$ is accordingly

$$O_2^{[p]}(z) = \left(\frac{1}{1-z}\right)^{2b-p+1} \times B_2^{[p]}(z).$$

Here, $B_2^{[p]}(z)$ is the generating function of the collection $\mathcal{B}_2^{[p]}$ and from our earlier discussion, it is a *polynomial* of degree at most $2\ell(\mathcal{D})$. Now, an easy dominant pole analysis entails that $[z^n]O_2^{[p]} = O(n^{2b-p})$.

This proves that the dominant contribution to the variance is given by $[z^n]O_2^{[1]}$, which is of order $O(n^{2b-1})$. Then, the variance $\mathbb{E}[\Xi_n^2]$ involves the constant $B_2^{[1]}(1)$ that is the total weight of the collection $\mathcal{B}_2^{[1]}$. Recall that this collection is formed of intersecting full pairs of occurrences of degree $2b-1$. The polynomial $B_2^{[1]}(z)$ is itself the generating function of the collection $\mathcal{B}_2^{[1]}$, and it is conceptually an extension of Guibas and Odlyzko's autocorrelation polynomial [20, 21]. We shall later make precise the relation between both polynomials (see Section 3.3).

We summarize our findings in the following theorem.

Theorem 1. *Consider a general constraint $\mathcal{D}$ with a number of blocks equal to b. The mean and the variance of the number of occurrences Ω of a pattern $\mathcal{W}$ subject to constraint $\mathcal{D}$ satisfy*

$$\begin{aligned} \mathbb{E}_n[\Omega] &= \frac{\pi(\mathcal{W})}{b!}\left(\prod_{j\,:\,d_j<\infty} d_j\right) n^b \left(1+O(\frac{1}{n})\right), \\ \mathbb{V}_n[\Omega] &= \sigma^2(\mathcal{W}) n^{2b-1}\left(1+O(\frac{1}{n})\right), \end{aligned}$$

where the "variance coefficient" $\sigma^2(\mathcal{W})$ involves the autocorrelation $\kappa(\mathcal{W})$

$$\sigma^2(\mathcal{W}) = \frac{\pi^2(\mathcal{W})}{(2b-1)!}\,\kappa^2(\mathcal{W}) \qquad \text{with} \quad \kappa^2(\mathcal{W}) := \sum_{(I,J)\in\mathcal{B}_2^{[1]}} \left(\frac{1}{\pi(\mathcal{W}_{I\cap J})} - 1\right). \tag{14}$$

The set $\mathcal{B}_2^{[1]}$ is the collection of all pairs of occurrences (I,J) that satisfy three conditions: (i) they are full; (ii) they are intersecting; (iii) there is a single pair (r,s) with $1 \le r,s \le b$ for which the rth *block $B^{[r]}$ of $\alpha(I)$ and the* sth *block $C^{[s]}$ of $\alpha(J)$ intersect.*

Remark. From Theorem 1 and Chebyshev's inequality we conclude that

$$\mathbb{P}_n\left\{\left|\frac{\Omega}{\mathbb{E}_n[\Omega]} - 1\right| > \epsilon\right\} \le \frac{\mathbb{V}_n[\Omega]}{\epsilon^2\,\mathbb{E}_n^2[\Omega]} = O\left(\frac{1}{n}\right).$$

Therefore, the random variables $\Omega/\,\mathbb{E}_n[\Omega]$ converge to 1 in probability, that is,

$$\text{for any } \epsilon > 0, \qquad \lim_{n\to\infty} \mathbb{P}_n\left\{\left|\frac{\Omega}{\mathbb{E}_n[\Omega]} - 1\right| < \epsilon\right\} = 1. \tag{15}$$

Bourdon and Vallée in [9] have recently provided a follow-up to Theorem 1 and shown that the statement is fairly robust: it extends to a somewhat larger class of patterns with gaps; more importantly perhaps, concentration of distribution is shown in [9] to hold for a wide class of sources encompassing memoryless and Markov sources—the *dynamical sources* in the sense of Vallée [10, 39].

3.3. **Generalized autocorrelations and variances.** In this subsection, we reexamine the variance coefficient, for which formulæ have been provided earlier; see (14) of Theorem 1. As we now explain, the variance coefficient turns out to be computable in a time that is polynomial in the size of the pattern specification. Structurally, it relates to a generalization of Guibas and Odlyzko's autocorrelation polynomial originally introduced for classical string matching (cf. [20, 21, 38]).

The general case. The computation of the autocorrelation $\kappa(\mathcal{W})$ reduces to b^2 computations of correlations $\kappa(\mathcal{W}_r, \mathcal{W}_s)$, relative to pairs $(\mathcal{W}_r, \mathcal{W}_s)$ of blocks. Note that each correlation of the form $\kappa(\mathcal{W}_r, \mathcal{W}_s)$ involves a totally constrained problem and is discussed below. Precisely, one has

$$(16)\quad \kappa^2(\mathcal{W}) = D^2(\mathcal{D}) \sum_{1\le r,s\le b} \frac{1}{D(\mathcal{D}_r)D(\mathcal{D}_s)} \binom{r+s-2}{r-1}\binom{2b-r-s}{b-r} \kappa(\mathcal{W}_r, \mathcal{W}_s),$$

where $\kappa(\mathcal{W}_r, \mathcal{W}_s)$ is the sum of the $e(I,J)$ taken over all full intersecting pairs (I,J) formed with an occurrence I of block $\mathcal{W}_r$ subject to constraint $\mathcal{D}_r$ and an occurrence J of block $\mathcal{W}_s$ subject to constraint $\mathcal{D}_s$. Let us explain the formula (16) in words: for a pair (I,J) of the set $\mathcal{B}_2^{[1]}$, there is a single pair (r,s) of indices with $1 \le r,s \le b$ for which the rth block $B^{[r]}$ of $\alpha(I)$ and the sth block $C^{[s]}$ of $\alpha(J)$ intersect. Then, there exist $r+s-2$ blocks before the block $\alpha(B^{[r]}, C^{[s]})$ and $2b-r-s$ blocks after it. We then have three different degrees of freedom: (i) the relative order of blocks $B^{[i]}(i<r)$ and blocks $C^{[j]}(j<s)$, and similarly the relative order of blocks $B^{[i]}(i>r)$ and blocks $C^{[j]}(j>s)$; (ii) the lengths of the blocks (there are D_j possible lengths for the jth block); (iii) finally the relative positions of the blocks $B^{[r]}$ and $C^{[s]}$.

The fully unconstrained case. In the unconstrained problem, the parameter b equals m, and each block $\mathcal{W}_r$ is reduced to the symbol w_r. Then the "correlation coefficient" $\kappa^2(\mathcal{W})$ simplifies to

$$(17)\qquad \kappa^2(\mathcal{W}) := \sum_{1\le r,s\le m} \binom{r+s-2}{r-1}\binom{2m-r-s}{m-r} [\![w_r = w_s]\!] \left(\frac{1}{p_{w_r}} - 1\right).$$

The totally constrained case. To complete the discussion relative to the variance coefficient, we need to show how to compute the correlation coefficient $\kappa(\mathcal{R}, \mathcal{S})$ between two totally constrained hidden patterns $(\mathcal{R}, \mathcal{C})$ and $(\mathcal{S}, \mathcal{D})$. (For general hidden patterns, $\mathcal{R}$ and $\mathcal{S}$ will be blocks of the original patern $\mathcal{W}$.) This is achieved by methods of dynamic programming. Assume that the 1–block pattern $\mathcal{R}$ has i symbols, so that constraint $\mathcal{C}$ is of the form $\mathcal{C} = (c_1, c_2, \ldots, c_{i-1})$; in the same vein, the 1–block pattern $\mathcal{S}$ has j symbols, so that constraint $\mathcal{D}$ is of the form $\mathcal{C} = (d_1, d_2, \ldots, d_{j-1})$.

When a pattern $\mathcal{T}$ occurs at a position I and $K \subset I$ is any subposition of I, $\mathcal{T}_K$ denotes the subpattern of $\mathcal{T}$ that occurs at position K. We consider the set $\mathcal{B}$ of pairs (I,J) that satisfy four conditions: (i) I is an occurrence of $\mathcal{R}$ with constraints $\mathcal{C}$; (ii) J is an occurrence of $\mathcal{S}$ with constraints $\mathcal{D}$; (iii) the two subpatterns $\mathcal{R}_{I\cap J}$ and $\mathcal{S}_{I\cap J}$ are equal; (iv) (I,J) is full. Now, the correlation coefficient $\kappa(\mathcal{R}, \mathcal{S})$ involves the set $\mathcal{B}$ is equal to

$$\kappa(\mathcal{R}, \mathcal{S}) := \sum_{(I,J)\in\mathcal{B}} e(I,J) \qquad \text{with} \quad e(I,J) = \frac{1}{\pi(\mathcal{R}_{I\cap J})} - 1.$$

The computation of $\kappa := \pi(\mathcal{R})\pi(\mathcal{S})\kappa(\mathcal{RS})$ reduces to that of A, C:

$$\kappa = A - \pi(\mathcal{R})\pi(\mathcal{S})C.$$

The quantities C, A are determined from auxiliary arrays Y, X:

$$C = \sum_{t,v=1}^{\ell(\mathcal{C})+\ell(\mathcal{D})} Y[t,v,i,j], \qquad A = \sum_{t,v=1}^{\ell(\mathcal{C})+\ell(\mathcal{D})} X[t,v,i,j]. \tag{18}$$

Arrays Y, X are themselves computed by recurrence with subsets $\mathcal{C}(t,k), \mathcal{D}(v,\ell)$,

$$\begin{array}{lcl} \mathcal{C}(t,k) & := & \{u; 1 \le u \le t-1, u \ge t - c_{k-1}\} \\ \mathcal{D}(v,\ell) & := & \{w; 1 \le w \le v-1, w \ge v - d_{\ell-1}\}. \end{array} \tag{19}$$

$$\begin{cases} Y[t,v,k,\ell] = \displaystyle\sum_{u \in \mathcal{C}(t,k)} Y[u,v,k-1,\ell] & \text{for } t > v \\ Y[t,v,k,\ell] = \displaystyle\sum_{w \in \mathcal{D}(v,\ell)} Y[t,w,k,\ell-1] & \text{for } t < v \\ Y[t,v,k,\ell] = [\![r_k = s_\ell]\!] \displaystyle\sum_{\substack{u \in \mathcal{C}(t,k) \\ w \in \mathcal{D}(v,\ell)}} Y[u,w,k-1,\ell-1] & \text{for } t = v \end{cases} \tag{20}$$

$$\begin{cases} X[t,v,k,\ell] = p(r_k) \displaystyle\sum_{u \in \mathcal{C}(t,k)} X[u,v,k-1,\ell] & \text{for } t > v \\ X[t,v,k,\ell] = p(s_\ell) \displaystyle\sum_{w \in \mathcal{D}(v,\ell)} X[t,w,k,\ell-1] & \text{for } t < v \\ X[t,v,k,\ell] = p(r_k)\, [\![r_k = s_\ell]\!] \displaystyle\sum_{\substack{u \in \mathcal{C}(t,k) \\ w \in \mathcal{D}(v,\ell)}} X[u,w,k-1,\ell-1] & \text{for } t = v \end{cases} \tag{21}$$

The initialization conditions are

$$\begin{array}{lll} Y[1,0,1,0] := 1; & Y[0,1,0,1] := 1; & Y[1,1,1,1] := [\![r_1 = s_1]\!]; \\ X[1,0,1,0] := p(r_1); & Y[0,1,0,1] := p(s_1); & X[1,1,1,1] := p(r_1)\, [\![r_1 = s_1]\!], \end{array} \tag{22}$$

and, for $(t,v) \neq (0,1)$ and $(t,v) \neq (1,0)$,

(23)

$$X[t,v,1,0] = Y[t,v,1,0] = X[t,v,0,1] = X[t,v,0,1] = Y[t,v,1,1] = X[t,v,1,1] = 0.$$

FIGURE 3. The formulæ summarizing the computation of the variance coefficient.

Remark that the condition for the pair (I, J) to be intersecting can be dispensed with, since non-intersecting pairs give rise to a term $e(I,J)$ equal to zero. An alternative expression of $\kappa := \pi(\mathcal{R})\pi(\mathcal{S})\kappa(\mathcal{R},\mathcal{S})$

$$\kappa = A - \pi(\mathcal{R})\pi(\mathcal{S})C \qquad \text{with} \qquad A = \sum_{(I,J)\in\mathcal{B}} \pi((\mathcal{R} \uparrow \mathcal{S})_{(I,J)}), \quad C = \sum_{(I,J)\in\mathcal{B}} 1$$

involves the pattern $(\mathcal{R} \uparrow \mathcal{S})_{(I,J)}$ obtained by merging the two patterns $\mathcal{R}$ at position I and $\mathcal{S}$ at position J. This merging is a sort of shuffle with possible collisions at position $I \cap J$.

From now on, the main formulæ of this subsection are grouped inside Fig. 3 which can be taken as an algorithm for determining variances. We consider the set $\mathcal{B}[t,v,k,\ell]$ of pairs (I,J) that satisfy the following: (i) I is a valid occurrence of prefix $\mathcal{R}_k$ whose last component i_k satisfies $i_k = t$, (ii) J is a valid occurrence of prefix $\mathcal{S}_\ell$ whose last component j_ℓ satisfies $j_\ell = v$, (iii) If I and J are intersecting, the equality $\mathcal{R}_{I\cap J} = \mathcal{S}_{I\cap J}$ holds, (iv) the pair (I, J) is full. Notice that the set $\mathcal{B}[t, t, k, \ell]$ is empty except if the last symbol r_k of $\mathcal{R}_k$ equals the last symbol s_ℓ of $\mathcal{S}_\ell$. Since the pair (I, J) is full, the indices t, v vary between 0 and $\ell(\mathcal{C}) + \ell(\mathcal{D})$. Index k is a cursor relative to pattern $\mathcal{R}$ (which varies between 0 and i), while index ℓ is a cursor inside pattern $\mathcal{S}$, (which varies between 0 and j). Two variables, $Y[t, v, k, \ell]$ and $X[t, v, k, \ell]$, are used. The first one represents the cardinality of the set $\mathcal{B}[t, v, k, \ell]$ and is used for computing the second term C of κ, while the second one is the total weight of this set and is used for computing the first term A of κ; see Equation (18) of Fig. 3.

The fundamental formulæ for $Y[t, v, k, \ell]$ and for $X[t, v, k, \ell]$ used for dynamic programming are of the same vein. For each of them, there appear three cases depending on the relative position of t and v (remark that equality $t = v$ is only possible if the equality $r_k = s_\ell$ holds). They both involve sets of indices defined from constraints $\mathcal{C}$ and $\mathcal{D}$ specified in Eq. (19) of Fig. 3, and auxiliary variables Y, X determined by the recurrences (20) and (21) respectively. (The formula for X is similar to that for Y, safe that it involves the probability of the last symbol read.) The variables must be initialized by (22). Moreover, since the pair (I, J) has to be full, except for $(t, v) = (1, 0)$ or $(t, v) = (0, 1)$, one sets the values as given in Fig. 3, Eq. (23).

The case of a string and the relation with autocorrelation polynomials. Here $\mathcal{W} = w_1 w_2 \dots w_m$ is a string of length m, and all the symbols of $\mathcal{W}$ must occur at consecutive places, so that a valid position I is an interval of length m. For $1 \le i \le j \le m$, we denote by $\mathcal{W}[i, j]$ the substring $w_i w_{i+1} \dots w_j$. The autocorrelation set $K_\mathcal{W} \subset [1..m]$ involves all indices k such that the prefix $\mathcal{W}[1, k]$ coincides with the suffix $\mathcal{W}[m - k + 1, m]$. Here, an index $k \in K_\mathcal{W}$ is relative to a intersecting pair of positions (I, J) and one has $\mathcal{W}[1\,..\,k] = \mathcal{W}_{I\cap J}$.

Classically, two autocorrelation polynomials, $A_\mathcal{W}$ and $C_\mathcal{W}$, are defined from $K_\mathcal{W}$. The polynomial $C_\mathcal{W}$ is the uniform autocorrelation polynomial while $A_\mathcal{W}$ is the weighted autocorrelation polynomial and involves suffix probabilities:

$$C_\mathcal{W}(z) = \sum_{k\in K_\mathcal{W}} z^{m-k},$$

$$A_\mathcal{W}(z) = \sum_{k\in K_\mathcal{W}} \pi(\mathcal{W}[k+1, m]) z^{m-k} = \pi(\mathcal{W}) \sum_{k\in K_\mathcal{W}} \frac{1}{\pi(\mathcal{W}[1, k])} z^{m-k}.$$

Since the polynomial $B_2^{[1]}$ involves coefficients of the form

$$\pi^2(\mathcal{W})[\frac{1}{\pi(\mathcal{W}_{I\cap J})} - 1],$$

this polynomial can be written as function of the two autocorrelations polynomials $A_\mathcal{W}$ and $C_\mathcal{W}$,

$$B_2^{[1]}(z) = \pi(\mathcal{W}) z^m \; [A_\mathcal{W}(z) - \pi(\mathcal{W})\, C_\mathcal{W}(z)].$$

Put simply, the variance coefficient of the hidden pattern problem extends the classical autocorrelation quantities associated with strings.

4. Central Limit Laws

Our goal is to prove that the sequence Ω_n appropriately centered and scaled tends to the normal distribution. We consider the following standardized random variable $\widetilde{\Xi}$ which is defined for each n by

$$\widetilde{\Xi}_n := \frac{\Xi_n}{n^{b-1/2}} = \frac{\Omega_n - \mathbb{E}_n[\Omega]}{n^{b-1/2}}, \tag{24}$$

where b is the number of blocks of the constraint $\mathcal{D}$. We shall show that $\widetilde{\Xi}$ behaves asymptotically as a normal variable with mean 0 and standard deviation σ. By the classical *moment convergence theorem* (Theorem 30.2 of [7]) this is established once all moments of $\widetilde{\Xi}_n$ are known to converge to the appropriate moments of the standard normal distribution. We remind the reader that if G is a standard normal variable (i.e., a Gaussian distributed variable with mean 0 and standard deviation 1), then for any integral $s \geq 0$

$$\mathbb{E}[G^{2s}] = 1 \cdot 3 \cdots (2s-1), \qquad \mathbb{E}[G^{2s+1}] = 0. \tag{25}$$

We shall accordingly distinguish two cases based on the parity of r, $r = 2s$ and $r = 2s+1$, and prove that

$$\mathbb{E}_n[\Xi^{2s+1}] = o(n^{(2s+1)(b-1/2)}), \qquad \mathbb{E}_n[\Xi^{2s}] \sim \sigma^{2s}\,(1 \cdot 3 \cdots (2s-1))\, n^{2sb-s}, \tag{26}$$

which implies Gaussian convergence of $\widetilde{\Xi}_n$.

Theorem 2. *The random variable Ω over a random text of length n asymptotically obeys a* Central Limit Law *in the sense that its distribution is asymptotically normal: for all $x = O(1)$, one has*

$$\lim_{n\to\infty} \mathbb{P}_n \left\{ \frac{\Omega - \mathbb{E}_n[\Omega]}{\sqrt{\mathbb{V}_n[\Omega]}} \leq x \right\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\, dt \tag{27}$$

Proof. The proof below is combinatorial; it basically reduces to grouping and enumerating adequately the various combinations of indices in the sum that expresses $\mathbb{E}_n[\Xi^r]$. Once more, $\mathcal{P}_n(\mathcal{D})$ is formed of all the sets of positions in $[1, n]$ subject to the constraint $\mathcal{D}$ and we set $\mathcal{P}(\mathcal{D}) := \bigcup_n \mathcal{P}_n(\mathcal{D})$. Then totally distributing the terms in Ξ^r yields

$$\mathbb{E}_n[\Xi^r] = \sum_{(I_1,\ldots,I_r) \in \mathcal{P}_n^r(\mathcal{D})} \mathbb{E}[Y_{I_1} \cdots Y_{I_r}]. \tag{28}$$

An r-tuple of sets $(I_1, \ldots, I_r)$ in $\mathcal{P}^r(\mathcal{D})$ is said to be *friendly* if each I_k intersects at least one other I_ℓ, with $\ell \neq k$ and we let $\mathcal{Q}^{(r)}(\mathcal{D})$ be the set of all friendly collections in $\mathcal{P}^r(\mathcal{D})$. For $\mathcal{P}^r$, $\mathcal{Q}^{(r)}$, and their derivatives below, we add the subscript n each time the situation is particularized to texts of length n. If $(I_1, \ldots, I_r)$ does not lie in $\mathcal{Q}^{(r)}(\mathcal{D})$, then $\mathbb{E}[Y_{I_1} \cdots Y_{I_r}] = 0$, since at least one of the Y_I's is independent of the other factors in the product and the Y_I's have been centered, $\mathbb{E}[Y_I] = 0$. One can thus restrict attention to friendly families and get the basic formula

$$\mathbb{E}_n[\Xi^r] = \sum_{(I_1,\ldots,I_r) \in \mathcal{Q}_n^{(r)}(\mathcal{D})} \mathbb{E}[Y_{I_1} \cdots Y_{I_r}], \tag{29}$$

where the expression involves fewer terms than in (28). From there, we proceed in two stages. First, restrict attention to friendly families that give rise to the dominant contribution and introduce a suitable subfamily $\mathcal{Q}_\star^{(r)} \subset \mathcal{Q}^{(r)}$; in so doing, moments of odd order appear to be negligible. Next, for even order r, the family $\mathcal{Q}_\star^{(r)}$ involves a symmetry and it suffices to consider another smaller subfamily $\mathcal{Q}_{\star\star}^{(r)} \subset \mathcal{Q}_\star^{(r)}$ that corresponds to a "standard" form of occurrence intersection; this last reduction precisely gives rise to the even Gaussian moments.

Odd moments. Given $(I_1, \dots, I_r) \in \mathcal{Q}^{(r)}$, the aggregate $\alpha(I_1, I_2, \dots, I_r)$ is defined as the aggregation (in the sense of the variance calculation above) of $\alpha(I_1) \cup \cdots \cup \alpha(I_r)$. Next, the *number of blocks* of $(I_1, \dots, I_r)$ is the number of blocks of the aggregate $\alpha(I_1, \dots, I_r)$; if p is the total number of intersecting blocks of the aggregate $\alpha(I_1, \dots, I_r)$, the aggregate $\alpha(I_1, I_2, \dots I_r)$ has $rb-p$ blocks. Like previously, we say that the family $(I_1, \dots, I_r)$ of $\mathcal{Q}_q^{(r)}$ is *full* if the aggregate $\alpha(I_1, I_2, \dots I_r)$ completely covers the interval $[1, q]$. In this case, the length of the aggregate is at most $rd(m-1)+1$, and the generating function of full families is a polynomial $P_r(z)$ of degree at most $rd(m-1)+1$ with $d = \max_{j \in \mathcal{F}} d_j$. Then, the generating function of families of $\mathcal{Q}^{(r)}$ whose block number equals k is of the form

$$\left(\frac{1}{1-z}\right)^{k+1} \times P_r(z),$$

so that the number of families of $\mathcal{Q}_n^{(r)}$ whose block number equals k is $O(n^k)$. This observation proves that the dominant contribution to (29) arises from friendly families with a maximal block number. It is clear that the minimum number of intersecting blocks of any element of $\mathcal{Q}^{(r)}$ equals $\lceil r/2 \rceil$, since it coincides exactly with the minimum number of edges of a graph with r vertices which contains no isolated vertex. Then the maximum block number of a friendly family equals $rb - \lceil r/2 \rceil$. In view of this fact and the remarks above regarding cardinalities, we immediately have

$$\mathbb{E}_n\left[\Xi^{2s+1}\right] = O\left(n^{(2s+1)b-s-1}\right) = o\left(n^{(2s+1)(b-1/2)}\right)$$

which establishes the limit form of odd moments in (26).

Even moments. We are thus left with estimating the even moments. The dominant term is relative to friendly families of $\mathcal{Q}^{(2s)}$ with an intersecting block number equal to s, whose set we denote by $\mathcal{Q}_\star^{(2s)}$. In such a family, each subset I_k intersects one and only one other subset I_ℓ. Furthermore, if the blocks of $\alpha(I_h)$ are denoted by $B_h^{[u]}, 1 \le u \le b$, there exists only one block $B_k^{[u_k]}$ of $\alpha(I_k)$ and only one block $B_\ell^{[u_\ell]}$ that contains the points of $I_k \cap I_\ell$. This defines an involution τ such that $\tau(k) = \ell$ and $\tau(\ell) = k$ for all pairs of indices (ℓ, k) for which I_k and I_ℓ intersect. Furthermore, given the symmetry relation $\mathbb{E}[Y_{I_1} \cdots Y_{I_{2s}}] = \mathbb{E}[Y_{I_{\rho(1)}} \cdots Y_{I_{\rho(2s)}}]$ it suffices to restrict attention to friendly families of $\mathcal{Q}_\star^{(2s)}$ for which the involution τ is the standard one with cycles $(1,2)$, $(3,4)$, etc; for such "standard" families whose set is denoted by $\mathcal{Q}_{\star\star}^{(2s)}$, the pairs that intersect are thus (I_1, I_2), $\dots$, (I_{2s-1}, I_{2s}). Since the set $\mathcal{K}_{2s}$ of involutions of $2s$ elements has cardinality $K_{2s} = 1 \cdot 3 \cdot 5 \cdots (2s-1)$ (cf. [16]), the equality

$$\sum_{\mathcal{Q}_{\star n}^{(2s)}} \mathbb{E}[Y_{I_1} \cdots Y_{I_{2s}}] = K_{2s} \sum_{\mathcal{Q}_{\star\star n}^{(2s)}} \mathbb{E}[Y_{I_1} \cdots Y_{I_{2s}}], \tag{30}$$

entails that we can work now solely with standard families.

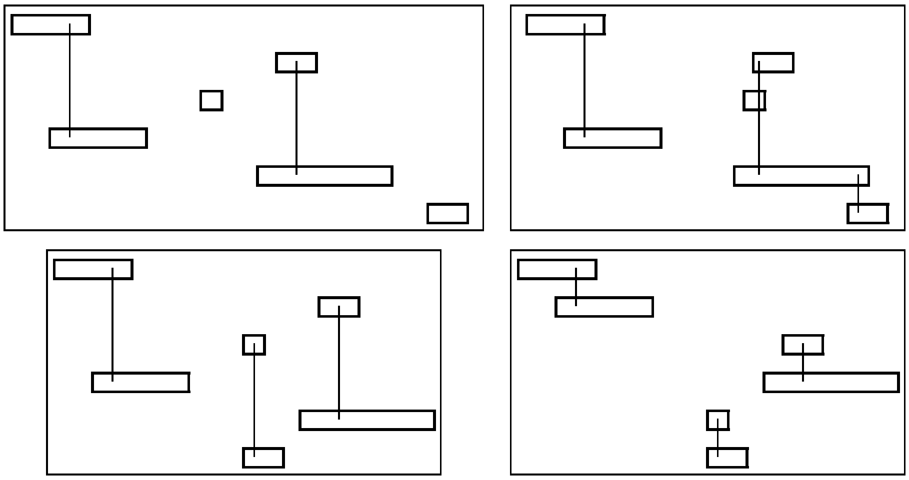


FIGURE 4. Various types of families of occurrence positions for $r = 2s = 6$: (i) an unfriendly family in $\mathcal{P}^r$; (ii) a friendly family in $\mathcal{Q}^{(r)}$; (iii) a friendly family with maximal number of joint blocks in $\mathcal{Q}_\star^{(2s)}$; (iv) a friendly family with maximal number of joint blocks and of standard type in $\mathcal{Q}_{\star\star}^{(2s)}$.

The class of occurrences relative to standard families is $\mathcal{A}^\star \times (\mathcal{A}^\star)^{2sb-s-1} \times \mathcal{B}_{2s}^{[s]} \times \mathcal{A}^\star$; this class involves the collection $\mathcal{B}_{2s}^{[s]}$ of all full friendly $2s$-tuples of occurrences with a number of blocks equal to s. Since $\mathcal{B}_{2s}^{[s]}$ is exactly a shuffle of s copies of $\mathcal{B}_2^{[1]}$ (as introduced in the study of the variance), the associated generating function is

$$\left(\frac{1}{1-z}\right)^{2sb-s+1} (2sb-s)! \left(\frac{B_2^{[1]}(z)}{(2b-1)!}\right)^s,$$

where $B_2^{[1]}(z)$ is the already introduced autocorrelation polynomial. Upon taking coefficients, we obtain the estimate

$$\sum_{\mathcal{Q}_{\star\star n}^{(2s)}} \mathbb{E}[Y_{I_1} \cdots Y_{I_{2s}}] \sim n^{(2b-1)s} \sigma^{2s}. \tag{31}$$

In view of the formulæ (28), (29), (30), and (31) above, this yields the estimate of even moments and leads to the second relation of (26). This completes the proof of Theorem 2. □

The even Gaussian moments eventually come out of the number of involutions, which corresponds to a fundamental asymptotic symmetry present in the problem. In this perspective the specialization of the proof to the fully unconstrained case is reminiscent of the derivation of the usual central limit theorem (dealing with sums of independent variables) by moments methods: compare with pp. 408–410 in Billingsley's book [7]. Proceeding along different tracks, Janson [26] has related this particular case to his treatment of U–statistics via Gaussian Hilbert spaces; see Chapter XI of Janson's book [25] for the type of method employed.

5. The Fully Constrained Case

This section develops the special case of a (fully) constrained pattern specified by a word $\mathcal{W} = w_1 w_2 \cdots w_m$ and the attached between-letters spacings corresponding to the constraint $\mathcal{D} = (d_1, d_2, \ldots, d_{m-1})$, where all the d_j's are finite. Like before, we set $D = \prod_j d_j$, and $\ell = \sum_j d_j$. The alphabet is $\mathcal{A} = \{a_1, \ldots, a_r\}$, and, in order to avoid trivialities, we assume its cardinality to be at least 2. Also, Ω denotes the parameter "number of occurrences of the pattern $(\mathcal{W}, \mathcal{D})$", so that, for some text T, the symbol $\Omega(T)$ denotes the number of occurrences of the pattern $(\mathcal{W}, \mathcal{D})$ in T. We can then use Ω to denote the corresponding random variable over the probability space $\mathcal{A}^n$ equipped with the Bernoulli (memoryless) model.

The mean and variance of Ω are, from earlier theorems, known to be of order $O(n)$. The central limit theorem is then applicable to this case. However, quite a bit more is available as expressed in the following statement:

Theorem 3. *Consider a fully constrained pattern with mean and variance coefficients $D\pi(\mathcal{W})$ and $\sigma^2(\mathcal{W})$.*

(i) *The random variable Ω satisfies a* Central Limit Law *with speed of convergence $1/\sqrt{n}$:*

$$\sup_x \left| \mathbb{P}_n \left\{ \frac{\Omega - D\pi(\mathcal{W})n}{\sigma(\mathcal{W})\sqrt{n}} \le x \right\} - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\, dt \right| = O\left(\frac{1}{\sqrt{n}}\right). \tag{32}$$

(ii) Large deviations *from the mean value have exponentially small probability: there exist a constant $\eta > 0$ and a nonnegative function I defined throughout $(0, \eta)$ such that $I(x) > 0$ for $x \neq D\pi(\mathcal{W})$ and*

$$\begin{cases} \displaystyle\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}_n \left(\frac{\Omega}{n} \le x \right) & = & -I(x) \quad \textit{if } 0 < x < D\pi(\mathcal{W}) \\ \displaystyle\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}_n \left(\frac{\Omega}{n} \ge x \right) & = & -I(x) \quad \textit{if } D\pi(\mathcal{W}) < x < \eta \end{cases}, \tag{33}$$

except for at most a finite number of exceptional values of x. Precisely, $I(x)$ can be computed as a function of an eigenvalue of a matrix (cf. Equation (48) below).

(iii) *Finally, for patterns called* primitive *(cf. Definition 1 below), a* Local Limit Law *holds:*

$$\sup_k \left| \mathbb{P}_n (\Omega = k) - \frac{1}{\sigma(\mathcal{W})\sqrt{n}} \frac{e^{x(k)^2/2}}{\sqrt{2\pi}} \right| = o\left(\frac{1}{\sqrt{n}}\right), \qquad x(k) = \frac{k - D\pi(\mathcal{W})n}{\sigma(\mathcal{W})\sqrt{n}}. \tag{34}$$

For patterns that are not primitive, it can be proved that, with d the period of the pattern, there exists a bounded quantity a (depending solely on the beginning and end of the text) such that

$$\widehat{\Omega} = \frac{\Omega - a}{d}$$

satisfies a local limit law in the sense of (34).

We proceed to establish this theorem in stages. First, we introduce a finite-state model: a deterministic finite automaton with weighted edges may be used to count all the occurrences of the pattern in texts. In fact, this automaton possesses a definite structure as it is a weighted variant of the classical de Bruijn graph. The finite-state property is then a reflection of the finiteness of all the gaps in the fully constrained case under study. This implies the existence of a matrix representation for our problem, a fact related to the technique of transfer matrices [6]; see Section 5.1). Then Perron-Frobenius properties and their perturbed versions apply, as detailed in Section 5.2; see especially Lemmas 2 and 3. A quasi-powers approximation (in the sense of Bender and Hwang [4, 23, 24]) for the probability generating function of Ω is then inferred, see Eq. (45). As developed in Section 5.3, this suffices to establish the central limit law (32) by a well-known process that parallels the usual proof of the central limit theorem for sums of independent random variables [4, 17, 23, 24]. Speed of convergence estimates expressed by (32) arise in this context from the Berry-Esseen inequalities. A similar analysis provides large deviation estimates as represented in a simplified form by (33). Additional strong positivity properties that are available when the pattern is *primitive* then induce estimates for the probabilities themselves (and not just for the cumulative distribution function), as expressed by (34).

Before embarking into technical developments, we briefly comment on the methodology employed in this section. The shape of our results is not unexpected since the central and local limit theorems that we obtain are closely related to matrix recursions developed in an important paper of Bender, Richmond, and Williamson [6]. The de Bruijn graph is classically associated with the combinatorial construction of de Bruijn sequences, and an early use of it in the context of word enumeration appears in [15]. Bender and Kochman in [5] make an implicit use of this construction combined with the central and local limit theorems of [6] to derive a very general class of estimates for subword counts. This shows the *shape* of the results that are to be expected in such situations, and our statement in Theorem 3 is definitely along these lines. However, the rather abstract character of the statements of [5] renders the specialization to our case somewhat unclear (to us at least), since a number of auxiliary technical conditions regarding nondegeneracy and aperiodicity would need to be established. For these reasons, we opt for a treatment that clearly draws its spirit from previous works [5, 6, 15, 23, 24], while remaining largely self-contained.

5.1. **The de Bruijn graph model.** First, we construct a matrix representation for the problem.

Lemma 1. *Consider a pattern $(\mathcal{W}, \mathcal{D})$, and let $\delta = \sum_j d_j = \ell(\mathcal{D}) - 1$ be the total length of all the gaps. Denote by Δ the quantity r^δ. There exist a matrix $T(u)$ of dimension $\Delta \times \Delta$ and two column vectors $\mathrm{x}(u)$, y of dimension Δ such that the probability generating function of the number of occurrences, Ω satisfies, for $n \geq \delta$,*

$$\mathbb{E}_n(u^\Omega) = \mathrm{x}(u)^{\mathbf{t}}\, T(u)^{n-\delta}\, \mathrm{y}.$$

The entries of $T(u)$ and $\mathrm{x}(u)$ are polynomials in u with nonnegative coefficients. The vector $\mathrm{y} = (1, \dots, 1)^{\mathbf{t}}$ is the column vector whose entries are all equal to the constant 1.

Proof. The basic idea amounts to constructing a device that scans the text $t_1 t_2 \cdots t_n$ and, at each stage, keeps in its (finite) memory the last δ letters read from the text. Formally, the de Bruijn graph is a finite automaton with state space $\mathcal{B} = \mathcal{A}^\delta$; the transition from a state $b \in \mathcal{B}$ upon scanning letter α is $\tau(b\alpha)$, where $\tau(f)$ for a word f just erases the leftmost symbol of f (this is a left shift of b concatenated

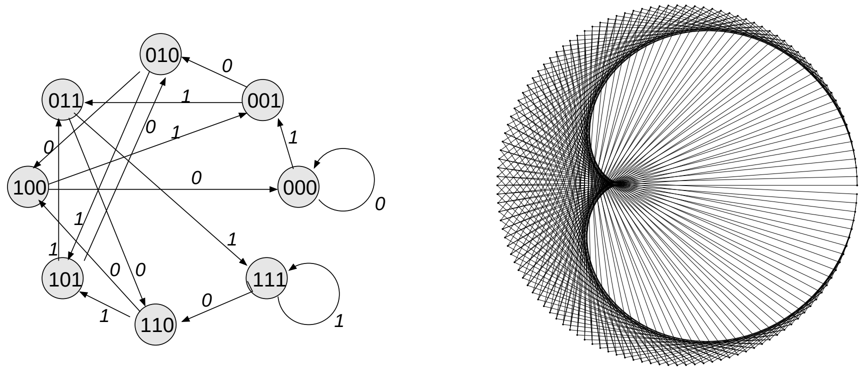


FIGURE 5. The de Bruijn graph corresponding to the binary alphabet $\mathcal{A} = \{0, 1\}$ and to block size equal to ℓ has 2^ℓ vertices (associated to blocks of length ℓ) and $2^{\ell+1}$ edges: the cases displayed are $\ell = 3$ (left) and $\ell = 7$ (right).

with α). A text of length $n \geq \delta$ is then associated to a path of length $n - \delta$ that begins at the state b formed with the first δ symbols of the text.

The de Bruijn graph lends itself to pleasant graphic renderings when vertices are ordered by lexicographic order and represented at regularly spaced points on a circle, with edges corresponding to nonzero entries in the transition matrix. Figure 5 exemplifies the case of a binary alphabet ($r = 2$) when the block size equals 3 or 7.

One can easily equip the automaton with a counter that gets incremented each time a transition is effected; this, in such a way that the value of the counter when the text is exhausted will contain the number Ω of occurrences of $\mathcal{W}$. Indeed, consider a transition $(b, \alpha) \mapsto c$ of the automaton; this requires $c = \tau(b\alpha)$ or equivalently $b\alpha \in \mathcal{A}c$. When this transition is effected, one can "cash in" all the "new" occurrences of $\mathcal{W}$ which arise when reading the last letter α, i.e., all the occurrences of the pattern that *end* at the letter α. Precisely, for a transition $(b, \alpha) \mapsto c$ of the automaton, the number of occurrences of the pattern $\mathcal{W}$ contained in $b\alpha$ and *ending* at the letter α is determined by either the pair (b, α) or the pair (b, c); we denote this number by $\phi(b, \alpha)$ or $\psi(b, c)$, depending on context, so that $\phi(b, \alpha) = \psi(b, c)$ whenever $c = \tau(b\alpha)$. Since the length of word $b\alpha$ exactly equals $\delta + 1 = \ell(\mathcal{W})$, all the occurrences of $\mathcal{W}$ that end at α are contained in a text of the form $b\alpha$ with $b \in \mathcal{A}^\delta$ so that the relation $\phi(b, \alpha) = \Omega(b\alpha) - \Omega(b)$ holds. We build a matrix $T(u)$ indexed by $\mathcal{B} \times \mathcal{B}$ as follows ($[\![\cdot]\!]$ is Iverson's bracket):

$$(35) \qquad (T(u))_{b,c} := p_\alpha u^{\phi(b,\alpha)} [\![\, b\alpha \in \mathcal{A}c \,]\!] = p_\alpha u^{\Omega(b\alpha) - \Omega(b)} [\![\, b\alpha \in \mathcal{A}c \,]\!].$$

By construction and by usual combinatorial properties of matrix products, the entry of index b, c of the power $T(u)^k$ cumulates all terms corresponding to starting in state b, ending in state c, and recording the total number of occurrences of the pattern $\mathcal{W}$ found upon scanning the last k letters of the text which allow the transition from state b to state c,

$$(36) \qquad \left(T(u)^k\right)_{b,c} = \sum_{f \in \mathcal{A}^k \ \mid \ bf \in \mathcal{A}^k c} \pi(f)\, u^{\Omega(bf) - \Omega(b)}.$$

Now, the entry of index b of the vector $\mathrm{x}(u)$ is simply to be taken as

$$(\mathrm{x}(u))_b = \pi(b)\, u^{\Omega(b)}.$$

Then, the summation of all the entries of the row vector $\mathrm{x}(u)^{\mathbf{t}} T(u)^k$ is achieved by means of the vector y so that the quantity $\mathrm{x}(u)^{\mathbf{t}} T(u)^k \mathrm{y}$ gives the probability generating function of Ω taken over all texts of length $\delta + k$. The statement follows upon setting $n = \delta + k$. ☐

Here is for instance the matrix associated with the pattern `a#₂b` corresponding to $(\mathcal{W}, \mathcal{D}) = (ab, 2)$, that is, occurrences of `ab` separated by at most one letter:

$$T(u) = \begin{array}{c} \\ aa \\ ab \\ ba \\ bb \end{array} \begin{array}{c} \begin{array}{cccc} aa & ab & ba & bb \end{array} \\ \begin{pmatrix} p_a & p_b u^2 & 0 & 0 \\ 0 & 0 & p_a & p_b u \\ p_a & p_b u & 0 & 0 \\ 0 & 0 & p_a & p_b \end{pmatrix} \end{array}. \tag{37}$$

5.2. **Spectral properties of transfer matrices.** Perron-Frobenius theory together with its analyticly perturbed versions provides valuable information on the growth of quantities attached to matrix powers. We develop here a basis of facts. In essence, the developments that follow are generic (i.e., applicable to any strongly connected graph equipped with a "flow" ψ defined on edges by integer weights). For the sake of simplicity, however, we only develop the theory on the particular instance of the de Bruijn graph.

Let $\lambda_j(u)$, for $j = 1\,..\,\Delta$ be a numbering of the eigenvalues of $T(u)$ taken so that $|\lambda_1(u)| \geq |\lambda_2(u)| \geq \cdots \geq |\lambda_\Delta(u)|$. The *spectral radius* of $T(u)$ is defined as usual as the maximum modulus of eigenvalues, $\rho(u) = |\lambda_1(u)|$. As it is well-known, the spectral radius governs the asymptotic growth of quantities involving $T(u)^n$, since, for any matrix norm $\|\cdot\|$, one has the property

$$\rho(u) = \lim_{n\to\infty} \|T(u)^n\|^{1/n}. \tag{38}$$

The following lemma summarizes some of the main properties of the dominant eigenvalue of $T(u)$ that intervene in the proof of Theorem 3.

Lemma 2. *Consider the matrix $T(u)$ relative to a fully constrained pattern $(\mathcal{W}, \mathcal{D})$. The following properties hold.*

(i) for $u > 0$, the matrix $T(u)$ has a unique dominant eigenvalue strictly positive denoted by $\lambda(u)$ and a dominant eigenvector $\mathrm{a}(u)$ whose entries are all strictly positive. There exists a complex neighborhood of the real positive axis on which the mappings $u \to \lambda(u), u \to \mathrm{a}(u)$ are well-defined and analytic; in addition, all the entries of $\mathrm{a}(u)$ are non zero.

(ii) At $u = 1$, the function $\lambda(u)$ satisfies

$$\lambda(1) = 1, \qquad \lambda'(1) = D\pi(\mathcal{W}), \qquad \lambda''(1) + \lambda'(1) - \lambda'(1)^2 = \sigma^2(\mathcal{W}). \tag{39}$$

For any cycle $\mathcal{C}$ of the de Bruijn graph, denote by $\psi(\mathcal{C}) := \sum_{(b,c)\in\mathcal{C}} \psi(b,c)$ the total weight of $\mathcal{C}$ relative to the pattern $(\mathcal{W}, \mathcal{D})$. One has also

$$\lim_{u\to 0^+} \frac{u\lambda'(u)}{\lambda(u)} = 0, \qquad \lim_{u\to+\infty} \frac{u\lambda'(u)}{\lambda(u)} = \eta := \max\left\{ \frac{\psi(\mathcal{C})}{|\mathcal{C}|}; \quad \mathcal{C} \text{ a cycle} \right\}. \tag{40}$$

(iii) For positive u, the function $u \to \lambda(u)$ is strictly increasing and its logarithm is strictly convex.

Formula (40) expresses information on the order of growth of $\lambda(u)$, namely, $\lambda(u) \asymp u^0$ (near 0^+) and $\lambda(u) \asymp u^\eta$ (near $+\infty$). Formulæ (39) and (40) are best understood when expressed in terms of the function $\Lambda(s)$, which has the character of a cumulant generating function:

$$\begin{cases} \Lambda(0) = 0, & \Lambda'(0) = D\pi(\mathcal{W}), \qquad \Lambda''(0) = \sigma^2(\mathcal{W}), \\ \lim_{s\to-\infty} \Lambda'(s) = 0, & \lim_{s\to+\infty} \Lambda'(s) = \eta. \end{cases}$$

(The Quasi-Powers approximation of (45) can be similarly interpreted in terms of cumulant generating function.)

Proof. (i) Take u real positive. Then, the matrix $T(u)$ has nonnegative entries, and for any exponent $L \geq \delta$, the Lth power of matrix $T(u)$ has strictly positive entries. This results from the fact that, for any $L \geq \delta$, there is always a path in the de Bruijn graph of length L connecting two states b and c; see also Equation (36). Then, the classical Perron-Frobenius theory of nonnegative matrices applies to matrix $T(u)$ (see, e.g., [18, Ch. 13]), to the effect that there exists an eigenvalue that dominates strictly all the other ones. Moreover this eigenvalue is simple and strictly positive. In other words, one has

$$\lambda(u) := \lambda_1(u) > |\lambda_2(u)| \geq |\lambda_3(u)| \geq \cdots . \tag{41}$$

as well as $\rho(u) = \lambda(u)$ for positive u. Also, by this theory, the eigenvector $\mathrm{a}(u)$ corresponding to $\lambda(u)$ has all its components that are strictly positive. Then, by classical (analytic) perturbation theory [27, Ch. II], there exists a neighbourhood of the real positive axis where the functions $u \to \lambda(u), u \to \mathrm{a}(u)$ remain well-defined and analytic in u. (In fact, $\lambda(u)$ is a branch of an algebraic function since it satisfies the characteristic equation $\det(\lambda I - T(u)) = 0$; an alternative direct proof of (i) could be given based on this observation.)

(ii) For $u = 1$, the matrix $T(u)$ is stochastic, so that $\lambda(1) = 1$. Two differentiations at $u = 1$ of the Quasi-Powers approximation, (45) below, show that the mean and variance of Ω_n are related to the first two derivatives of $\lambda(u)$ at 1. This establishes the relations (39).

We next prove the relation (40). We shall only do so for the maximum, describing the behaviour of $\lambda(u)$ as $u \to \infty$, since the dual relation at 0 follows from similar arguments (based on the minimum of the ψ values along cycles). Let $p_n(u)$ denote $\mathbb{E}_n(u^\Omega)$. The maximum relation in (40) is equivalent to asserting the coincidence of the combinatorially defined "cyclic index" η of the flow ψ with an analytically defined "Puiseux index" ϖ, as we now explain.

The *cyclic index* is defined as

$$\eta := \max_{c\,:\,\text{simple cycle}} \frac{\psi(c)}{|c|}.$$

Any path in a graph decomposes into a short header, a short trailer, and a collection of simple cycles—the construction is akin to the loop-erasing random walk. As a consequence, the cyclic index is seen to satisfy

$$\eta = \sup_{c\text{ path}} \frac{\psi(p)}{|p|} = \sup_{c\text{ cycle}} \frac{\psi(c)}{|c|}. \tag{42}$$

In other words, the cyclic index determines the worst-case behaviour of ψ on long paths.

The Puiseux index is defined as

$$\varpi := \lim_{u\to\infty} \frac{\log \lambda(u)}{\log u},$$

its existence being guaranteed by general properties of algebraic functions (the Newton–Puiseux theorem). The coincidence of ϖ and η then derives from the fact that there exist two positive constants A, B such that, for $n \geq n_0$ and all $u > 1$, one has

$$u^{-C_3} A^n u^{\eta n} < p_n(u) < u^{C_2} B^n u^{\eta n}, \tag{43}$$

where C_2, C_3 are (unessential) constants. Indeed, taking nth roots in (43) and passing to the limit, one finds for any $u > 1$,

$$Au^\eta \leq \lambda(u) \leq Bu^\eta,$$

an inequality that is incompatible with $\varpi \neq \eta$.

There only remains to justify (43). For the upper bound, observe that the ψ-value of any path of length n is at most $\eta\, n + C_2$ (for some C_2) by previous considerations, while the total number of paths of length n is bounded from above by Δ^{n+1} and the probability of any such path is at most P^n, where P is the largest of all the edge probabilities. For the lower bound, observe that there is at least a path of length n having weight $n\,\eta + O(1)$ (obtained by repeating a maximal simple cycle), this path having probability at least $\overline{P}^n$, with $\overline{P}$ the smallest of all edge probabilities.

(iii) The increasing property for $\lambda(u)$ depends on the well-known fact that if A and B are nonnegative irreducible matrices such that $A_{i,j} \geq B_{i,j}$ for all (i, j), then the spectral radius of A is larger than the spectral radius of B. (This easily results from the matrix norm property (38).)

By a well-known property, any (nondegenerate) probability generating function $f(u)$ is strictly log-convex at positive points within its domain of convergence, namely

$$\log f\left(\frac{a+b}{2}\right) < \frac{1}{2}\left(\log f(a) + \log f(b)\right),$$

for $a \neq b$. This relation is *a fortiori* valid for the probability generating function $p_n(u) = \mathbb{E}_n(u^\Omega)$ given by Lemma 1, which satisfies the Quasi-Powers approximation of (45) below. Taking nth root and passing to limit shows convexity properties to be inherited by $\lambda(u)$. $\square$

For the continuation of our analytic treatment, the notions of primitivity and period must be introduced.

Definition 1. *Let again $\psi(\mathcal{C})$ be the total weight of cycle $\mathcal{C}$ in the de Bruijn graph relative to the pattern $(\mathcal{W}, \mathcal{D})$. The quantity $\psi_{(\mathcal{W},\mathcal{D})} := \gcd\{\psi(\mathcal{C});\ \mathcal{C}$ a cycle$\}$ is called the* period. *Accordingly, a pattern is said to be* primitive *when its period is equal to 1.*

Lemma 3. *The following additional properties of the spectral radius of $T(u)$ hold.*

- (iv) *For any $\theta \in]0, 2\pi[$, one has $\rho(re^{i\theta}) \leq \rho(r)$.*
- (v) *Let $d = \psi_{(\mathcal{W},\mathcal{D})}$ be the period of pattern $(\mathcal{W}, \mathcal{D})$.*
  - $(v.a)$ *When $d = 1$, then $\rho(re^{i\theta}) < \rho(r)$ for all $\theta \in]0, 2\pi[$.*
  - $(v.b)$ *When $d > 1$, then $\rho(re^{i\theta}) = \rho(r)$ if and only if $\theta = 2k\pi/d$. In this case, the characteristic polynomial* $\det(\lambda I - T(u))$ *of matrix $T(u)$ is a polynomial of $\mathbb{R}[u^d, \lambda]$.*

Proof. We denote by $p_{i|j}$ the probability of the transition $j \to i$.

(iv) This part easily results from domination properties of matrices (cf. the argument used for Part (iii) of Lemma 2). However, as a preparation to the later part of the proof, we offer an alternative argument. For $|u| = 1$, and r real positive, consider the two matrices ${}^{\mathbf{t}}T(r)$ and ${}^{\mathbf{t}}T(ru)$. With (i), there exist a dominant eigenvalue $\lambda := \lambda(r)$ strictly positive and a dominant eigenvector $\mathrm{a} := \mathrm{a}(r)$ of ${}^{\mathbf{t}}T(r)$ relative to $\lambda(r)$ whose all entries a_j are strictly positive. Consider an eigenvalue μ of ${}^{\mathbf{t}}T(ru)$ and an eigenvector c relative to μ. Denote by v_j the ratio c_j/a_j. One can always choose vectors a and c such that $\max_{1\le j\le\Delta}|v_j| = 1$. Suppose that this maximum is attained for some index i. One has

$$|\mu c_i| = |\sum_j p_{i|j}\,(ru)^{\psi(j,i)}\,c_j| \le \sum_j p_{i|j}\,r^{\psi(j,i)}\,a_j = \lambda a_i, \tag{44}$$

so that $|\mu| \le \lambda$, and (iv) is established.

(v) Suppose now that the equality $|\mu| = \lambda$ holds. Then, the previous inequalities (44) all become equalities. First, for all indices ℓ such that $p_{i|\ell} \ne 0$, we deduce that $|c_\ell| = a_\ell$, so that v_ℓ has modulus 1. For these indices ℓ, we have the same equalities in (44) as previously for i. Finally, the transitivity of the de Bruijn graph entails that that each complex v_j is of modulus 1. Now, the converse of the triangular inequality shows the relation,

$$\text{for each edge } (j,i),\;\; u^{\psi(j,i)}v_j = \frac{\mu}{\lambda}v_i,$$

so that,

$$\text{for any cycle of length } L,\; \left(\frac{\mu}{\lambda}\right)^L = u^{\psi(\mathcal{C})}.$$

However, for any pattern $\mathcal{W}$, there exists a cycle $\mathcal{C}$ of length one with weight $\psi(\mathcal{C}) = 0$: if $\beta \in \mathcal{A}$ is distinct from the last symbol w_m of $\mathcal{W}$, the cycle labelled by β that starts at β^δ is convenient. This proves that $\mu = \lambda$ and that $u^{\psi(\mathcal{C})} = 1$ for any cycle $\mathcal{C}$.

Denote by $\psi_{\mathcal{W}}$ the gcd of all the quantities $\psi(\mathcal{C})$. If the period of $(\mathcal{W}, \mathcal{D})$ equals 1, $\psi_{\mathcal{W}} = 1$, then $u = 1$ and $(v.a)$ is proven.

As regards $(v.b)$, suppose now that the period is some integer $d > 1$. Then, for any integer k, the trace of the matrix $T(v)^k$ is a polynomial in v^d, so that the characteristic polynomial whose coefficients can all be expressed with these traces belongs to $\mathbb{R}[v^d, z]$. Consequently, the dominant eigenvalue $\lambda(v)$ is itself a function of v^d. □

Observe, as a consequence of the discussion of Part (v), that the period d is effectively computable via the symbolic form of the characteristic polynomial of matrix $T(u)$.

We conclude by listing situations where the hidden pattern $\mathcal{W}$ is guaranteed to be primitive. The conditions given in the following statement are likely to cover most cases of practical interest, although a few patterns will be left out like, in the Latin alphabet, "*The quick brown foxes jump over lazy dogs*"(!). (It might even be the case that all patterns are primitive, but we do not have a proof of this fact.)

Lemma 4. *The following are sufficient conditions for a pattern to be primitive:*

- (a) *$\mathcal{W}$ is a string, that is, all spacings satisfy $d_j = 1$;*
- (b) *the pattern alphabet is incomplete: at least one symbol of the alphabet $\mathcal{A}$ does not appear in $\mathcal{W}$;*

(c) *the symbols $w_1, w_2, \ldots, w_{m-1}$ each differ from the last symbol w_m.*

Proof. (a) Consider first a string $\mathcal{W}$, and denote by i the first index $i > 0$ where the autocorrelation polynomial has a non zero coefficient c_i. If there does not exist such an index, let $i := m$. Then the cycle that starts at the state $w_1 w_2 \ldots w_{m-1}$ and whose successive edges are labelled by the i symbols $w_{m-i}, w_{m-i+1}, \ldots w_{m-1}$ of $\mathcal{W}$ has a total weight equal to one, since the first edge has a weight equal to 1 while all the other edges have a zero weight.

Consider finally a pattern $(\mathcal{W}, \mathcal{D})$ which is not a string. Since it possesses at least one gap at least equal to 2, one has $m \leq \delta$. Let $\delta = m + p$,.

(b) Choose a letter $z \in \mathcal{A}$ not occurring in the pattern. Consider the cycle that starts at state $b := z^p \mathcal{W}$, and whose edges are labelled by successive symbols of $z^{\delta+1}\mathcal{W}$. Clearly this cycle has a weight equal to 1.

(c) Suppose now that all the symbols $w_1, w_2, \ldots w_{m-1}$ differ from the last symbol w_m. Choose a letter $z \in \mathcal{A}$ distinct of w_1 and w_m. Consider the cycle that starts at state $b := z^p \mathcal{W}$, and whose edges are labelled by successive symbols of b. Clearly this cycle has a weight equal to 1.

□

5.3. **Distributional properties.** We now apply the results derived in the previous subsections to fine characterizations of the law of the number of occurrences, that is, we complete the proof of Theorem 3.

As already remarked, the spectral radius and the dominant eigenvalue dictate the growth of all the entries of matrix powers $T(u)^n$. Then, by Lemma 2 (i), for u on or near the positive real line, the matrix $T(u)$ has a dominant eigenvalue $\lambda(u)$ which is unique, and strictly dominates all the other eigenvalues. Consequently, there exists a constant $A < 1$ such that $|\lambda_2(u)|/\lambda(u) < A < 1$. More precisely, the spectral decomposition of $T(u)$ when u lies in a sufficiently small complex neighbourhhod of any compact subinterval of $(0, +\infty)$ is of the form

$$T(u) = \lambda(u)Q(u) + R(u)$$

where $Q(u)$ is the projection under the dominant eigensubspace and $R(u)$ a matrix whose spectral radius equals $|\lambda_2(u)|$. Now, for any $n \geq 1$, the decomposition

$$T(u)^n = \lambda(u)^n Q(u) + R(u)^n,$$

entails, with Lemma 1 granting $\mathbb{E}_n[u^\Omega] = \mathrm{x}(u)^{\mathbf{t}} T(u)^{n-\delta}\mathrm{y}$, the estimate

$$\mathbb{E}_n[u^\Omega] = c(u)\lambda(u)^{n-\delta}\left(1 + O(A^n)\right), \tag{45}$$

for a nonzero analytic function $c(u)$. A uniform approximation like (45) for a sequence of probability generating functions is known as a *Quasi-Powers approximation.* Its existence in regions around 1, $+\infty$, and on the unit circle are respectively associated with central limits, large deviations, and local limits [4, 23, 24], as we see now.

Central limit law. Given a Quasi-Powers approximation valid when u lies in a complex neighbourhhod of 1, the classical proof of the central limit theorem for sums of independent random variables [19] can be mimicked and convergence to the Gaussian distribution results, following Bender and Hwang [4, 24]. The speed of convergence is found to be $O(1/\sqrt{n})$ as results from the Berry-Esseen inequalities; see [24] for the general argument. In this way, Eq. (32) of Theorem 3 is established.

Large deviations. We next consider large deviations and relate them to the existence of a Quasi-Powers approximation along the positive real axis. We concentrate on the left part of the distribution and write $p_{n,k} = \mathbb{P}_n(\Omega = k)$, $p_n(u) = \mathbb{E}_n(u^\Omega)$. First, by trivial bounds, one has elementarily

$$\sum_{k \le xn} p_{n,k} = [u^{\lfloor xn \rfloor}] \frac{p_n(u)}{1-u} \le \frac{p_n(\theta)}{(1-\theta)\theta^{\lfloor xn \rfloor}}, \tag{46}$$

for any fixed $\theta \in (0,1)$. Then, the Quasi-Powers approximation (45) applied to (46) yields

$$\sum_{k \le xn} p_{n,k} = O\left(\frac{\lambda(\theta)^n}{\theta^{xn}}\right). \tag{47}$$

The next move consists in adopting in (47) the particular value of θ that produces the best upper bound in (47). To this effect, define

$$I(x) = -\log \frac{\lambda(\zeta)}{\zeta^x} \qquad \text{with } \zeta \equiv \zeta(x) \in (0,1) \text{ defined by } \quad \frac{\zeta\lambda'(\zeta)}{\lambda(\zeta)} = x. \tag{48}$$

(The existence of ζ is guaranteed by (ii) and (iii) of Lemma 2.) Then, the upper bound (47) becomes

$$\frac{1}{n} \log \mathbb{P}_n\left(\frac{\Omega}{n} \le x\right) \le -I(x) + o(1). \tag{49}$$

There remains to prove that the upper bound coincides with the right rate. This is done following a classical technique of Cramér, also known as "shifting the mean". To wit, introduce the shifted version Y_ζ of Ω defined by

$$Y_\zeta : \quad \mathbb{E}(u^{Y_\zeta}) = \frac{p_n(\zeta u)}{p_n(\zeta)},$$

for the particular ζ of (48). The shifted Y_ζ satisfies a Quasi-Powers approximation in the central region $u \approx 1$ and is thus asymptotically normal provided

$$\lim_{n\to\infty} \frac{1}{n} \mathbb{V}(Y_\theta) = \lambda''(\zeta) + \lambda'(\zeta) - \lambda'(\zeta)^2 \tag{50}$$

is nonzero. The limit quantity in (50) represents an analytic function of ζ that is nonzero at $\zeta = 1$ and hence can only vanish sporadically at most at a finite set of isolated points. Except possibly for such isolated values, the quantity

$$\frac{1}{p_n(\zeta)} \sum_{xn-\sqrt{n}<j\le xn} p_{n,k}\zeta^j \tag{51}$$

then tends to a nonzero limit (expressible as a Gaussian error function). Since the weights in (51) are all of the form $\zeta^{xn}e^{O(\sqrt{n})}$, a lower bound on the $p_{n,k}$ follows. Thanks to this, the inequality in (49) can then be changed to equality, which is what Theorem 3 asserts.

A mirror argument (with ζ taken larger than 1) establishes the right part of the large deviation estimate in Theorem 3. Observe that Conditions (ii) and (iii) of Lemma 2 guarantee the existence of a suitable value of ζ over the *complete range* of the distribution of Ω_n.

Local limit law. Stronger "regularity conditions" are needed in order to obtain local limit estimates. Roughly, one wants to exclude the possibility that the discrete distribution is of lattice type, being supported by a nontrivial sublattice of the integers. (For instance, we need to exclude the possibility for Ω to be always odd, or of the parity of n, and so on.) Observe first that positivity and irreducibility of the matrix $T(u)$ are not enough. For instance the matrix

$$M = \begin{pmatrix} 1 & u^4 \\ u^2 & u^3 \end{pmatrix}$$

has a spectrum that depends on u via u^3 only. In particular, the spectral radius is a function of u^3. It is precisely this type of pathological behaviour that is excluded in the case when $T(u)$ stems from a primitive pattern.

Granted Lemma 3, one can estimate the probability distribution of Ω by the classical saddle point method in the case when $\mathcal{W}$ is primitive. This is similar to what is done to establish local limit laws for sums of discrete random variables [19]. One starts from Cauchy's coefficient integral,

$$\mathbb{P}_n(\Omega = k) = \frac{1}{2i\pi} \int_{|z|=1} p_n(u) \, \frac{du}{u^{k+1}}, \tag{52}$$

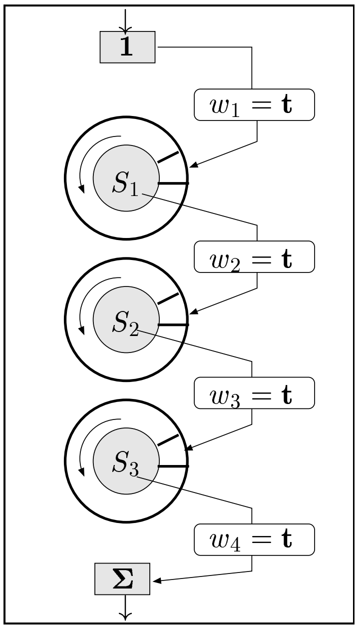
where k is now of the form $k = D\pi(\mathcal{W})n + x\sigma(\mathcal{W})\sqrt{n}$. Property $(v.a)$ of Lemma 3 grants us precisely the fact that any closed arc of the unit circle not containing $z = 1$ brings an exponentially negligible contribution. A standard application of the saddle point technique (details omitted) does the job. In this way, the proof of the local limit law, Eq. (34) of Theorem 3 is completed.

Theorem 3 invitingly points to similar statements that would be applicable to general hidden patterns. Guivarc'h (personal communication) has suggested the use of the theory of random walks on nilpotent Lie groups, as the pattern counting problem can be expressed as a product of random matrices that are nilpotent deformations of the identity; see [22] for a survey of some of the relevant methods. Also, Janson [26] has very recently obtained bounds for large deviations in the general case of hidden-words statistics using a generalization of Hoeffding's method for dependent random variables.

6. Experiments

It is of interest to try and assess the relevance of our analyses in contexts closer to real-life applications. For this purpose, we have set up a small campaign of experiments on "actual" data, in fact, pieces of English text. These experiments have no pretense of constituting an exhaustive study. They are merely intended as a coarse verification of some of the major phenomena inherent in hidden-pattern matching. Since the source model considered, of the memoryless type, is rather simplistic, one could be fairly satisfied with analytical results that correctly predict at least the orders of magnitude of the observed phenomena.

The experiment have been conducted with our own dynamic programming implementation of (constrained and unconstrained) sequence comparison and start with a brief discussion of the algorithmic complexity issues involved. Globally, both the "recognition problem" (i.e., does a pattern occur or not?) and the "reporting problem" (i.e., report the number of all occurrences and possibly a factored representation of the occurrence places) may be considered. In the unconstrained case, the recognition problem can be solved simply by a deterministic finite automaton (DFA) with m states, so that its complexity is $O(n)$. For the reporting problem,



The machine corresponding to a single block ($b = 1$) is composed of wheels and of an output counter Σ as depicted on the left (the number of blocks is $b = 1$ and the pattern length is $m = 4$).
The wheel of index j is divided into d_j sectors that each keep the number of valid occurrences of $w_1, \ldots, w_j$ at distance $0, 1, \ldots, (d_j - 1)$ in the past (the ones that have not yet expired); the attached quantity S_j is the sum of the values currently contained in all the sectors of wheel j.
When a new character $\mathbf{t}$ is read from the text, the active sectors are updated: if $\mathbf{t} \neq w_j$, then the active sector of wheel j is set to 0; if $\mathbf{t} = w_j$, then this sector is set to S_{j-1}. The running sum S_j is updated accordingly. Then the wheels are all rotated.
The top counter provides a continuous source of $\mathbf{1}$'s. The bottom counter Σ provides the number of occurrences of the constrained pattern seen so far: it maintains the cumulated sum of all the values passed by wheel $m - 1$, that is, of the S_{m-1}'s.
The case where the number of blocks satisfies $b > 1$ is obtained by stacking b elementary machines corresponding to the individual blocks.

FIGURE 6. A hidden-pattern counting machine.

the basic dynamic programming algorithm has cost $O(nm)$. (This is a simplification of the Longest Common Subsequence algorithm.) In the constrained case, a DFA can be set up so that the complexity of the recognition problem is $O(n)$, but the preprocessing and storage costs are exponential in the size of the pattern specification, which is certainly prohibitive in most practical application. Alternatives exist: see, e.g., [30] for a flavour of the methods (Directed Acyclic Word Graphs and suffix trees are useful) and Kucherov's fast implementation called `grappe` of the recognition problem. With d being the maximum gap allowed between letters, the reporting problem can be solved by a suitable implementation of the dynamic programming approach in total time $O((n + d)m)$—this is a simple programming exercise in circular list management; see Figure 6.

We used a piece of natural language text both as a source of characters and as a source of words. The complete works of Shakespeare are found under

`http://the-tech.mit.edu/Shakespeare/.`

We first extracted the full text of Hamlet stripped of all the comments:

> **Hamlet:** Who's there? || Nay, answer me: stand, and unfold yourself. || Long live the king! || Bernardo? || He. || You come most carefully upon your hour. [...]

In this, all nonalphabetic characters are suppressed and upper-case letters are normalized to lower case. This gives us a (rather unpoetical looking) text that has one long line with 150,372 characters:

		w = thelawisgaussian		$\widetilde{w}$ = naissuagsiwaleht	
d	Expected (E)	Occurred (Ω)	Ω/E	Occurred (Ω)	Ω/E
13	9.195E+01	0	0.00	18	0.19
14	2.794E+02	693	2.47	371	1.32
15	7.866E+02	1,526	5.46	2,379	3.02
18	1.211E+04	31,385	2.58	14,123	1.16
20	5.886E+04	124,499	2.11	41,066	0.69
25	1.673E+06	2,527,148	1.51	1,277,584	0.76
30	2.577E+07	40,001,940	1.55	25,631,589	0.99
40	1.928E+09	2,757,171,648	1.42	2,144,491,367	1.11
50	5.482E+10	76,146,232,395	1.38	48,386,404,680	0.88
∞	1.330E+48	1.36554E+48	1.03	1.38807E+48	1.04

FIGURE 7. Observed occurrences (Ω) versus predicted values (expectations, E) in the alphabetical characters of Hamlet.

$\mathbf{H_0}$: who s there nay answer me stand and unfold yourself long live [...]

Stripped of its spaces (‘ ’), the text now shrinks to $n = 120,057$ characters:

$\mathbf{H_1}$: whostherenayanswermestandandunfoldyourselflonglive [...]

This text, H_1, is the one used for experiments.

As somewhat arbitrary patterns, we adopt the phrase, "*The law is Gaussian*", and its mirror image,

$$\mathcal{W}_0 = \textsf{thelawisgaussian}, \qquad \widetilde{\mathcal{W}_0} = \textsf{naissuagsiwaleht},$$

corresponding to $m = 16$. Consider first the (fully) unconstrained case. If letters were all equally likely the configuration $n = 120,057$, $m = 16$ and alphabet cardinality $r = 26$ would lead us to expect a number of occurrences of $\mathcal{W}_0$ or $\widetilde{\mathcal{W}_0}$ about $5 \cdot 10^{87}$. The observed counts, which are $1.365\,10^{48}$ and $1.388\,10^{48}$ respectively, are *much* smaller. In fact, when estimated from the empirical distribution of letter frequencies in the text, the expected number of occurrences drops to $1.330\,10^{48}$, so that the observed counts only deviate by less than 5% from what is expected. Turning to the (fully) constrained problem, say we bound uniformly the separation distance between any two letters by d. Analysis (based on the natural frequencies of letters in the text) predicts that the pattern might start occurring near $d = 10$, while its presence is unlikely for smaller values, $d < 10$. In the text, w starts occurring at $d = 14$ while $\widetilde{w}$ starts at $d = 13$—a deviation of some 30–40% from what the model predicts. A table of observed versus predicted values when d varies is given in Figure 7. This shows a fair fit between the theoretical model and the observed data even though the text chosen is far from being "random" (and memoryless!).

Globally, as is perceptible from Figure 7, the less constrained patterns (d large or even $d = \infty$) are the ones in closest agreement with theory. Indeed, the fact that sequences like "the" or "law"are naturally present in English seems to give an advantage to pattern $\mathcal{W}$ for small values of d. (For instance, based on letter frequencies, the string the would be expected to occur 85 times but is actually present 1972 times in the text.) In contrast, the mirror image $\widetilde{\mathcal{W}}$, which has no clear "natural" structure, tends to be more compliant to theory. (Such fine phenomena would most likely be well captured by a Markovian model; see [9] for such an extension.) Figure 8 further illustrates this by displaying the evolution of the ratios Observed/Expected (Ω/E) as letters in the text are scanned one by one.

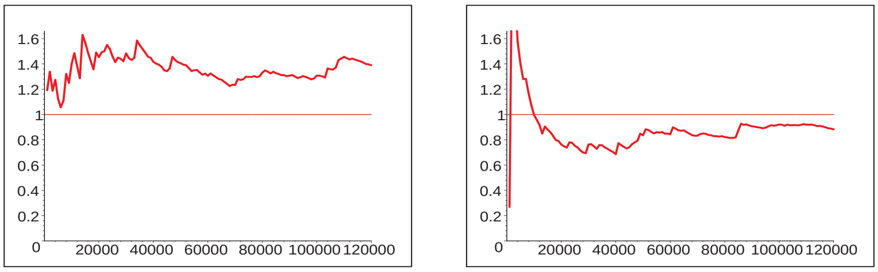


FIGURE 8. The evolutions of Ω/E for the 1 block patterns $\mathcal{W}$ (left) and $\widetilde{\mathcal{W}}$ (right) when the separation distance is $d = 50$.

As yet another test, we have examined the evolution of occurrence ratios Ω/E for two patterns, namely

$$\mathcal{W}_1 = \mathsf{ffffff}\ (d = 25), \qquad \mathcal{W}_2 = \mathsf{iamtheking}\ (d = 50).$$

These are displayed by thick lines in Figure 9. The "advantage" of pattern $\mathcal{W}_2$ which is in the language is perceptible as the number of observed occurrences is about twice what is expected. For comparison, we have also plotted the similar evolutions, but now relative to 5 random permutations of the Hamlet text (dashed lines): this conveys an impression of the ambient stochastic fluctuations, but also shows at the same time that, for random text, both patterns conform comparably well with what theory predicts.

The data so far have concerned events (letters) corresponding to characters in the text. We next turn to a situation where elementary events are words (now playing the rôle of individual letters) and a pattern is a succession of events satisfying various distance constraints. There is however a statistical difficulty as most meaningful words have a rather small probability of occurrence, so that any reasonably complex pattern is almost surely not observed at all. The text of Hamlet comprises 30,316 words, of which 4490 are different (with related forms like close, closely, closes, closet, or command, commanded, commandment, commands). In view of possible data mining applications, where it is mostly rough "contents" (roots?)

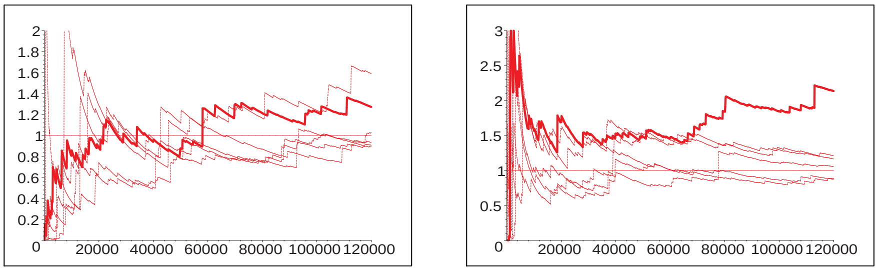


FIGURE 9. The evolutions of Ω/E for patterns $\mathcal{W}_1$ (left) and $\mathcal{W}_2$ (right): a comparison between the original text (thick lines) and 5 randomly permuted versions (dashed lines).

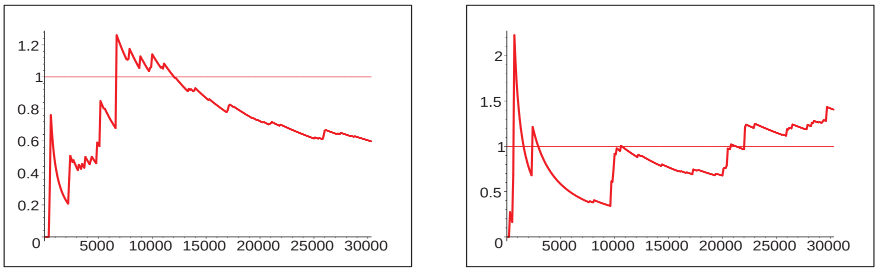


FIGURE 10. The Soundex-ed version of Hamlet. The patterns are $\mathcal{W}_3 \equiv$ "*to be or not to be*" and $\mathcal{W}_5 \equiv$ "*who is the king*" taken with distance $d = 100$.

of words that matter, we simplify the text of Hamlet by applying the Soundex algorithm. (The Soundex algorithm as described by Knuth in [28] is intended to hash words (in particular surnames) into a small space using a simple model which approximates the sound of the word when spoken by an English speaker. For instance, Gauss and Gosh both hash to G200; Hilbert and Heilbronn to H416.) When subjected to this transformation, *Hamlet* consists of "letters" in the form of compressed words (each formed of four alphanumerical characters). Under this encoding, the text of Hamlet (H_0) becomes the even less poetical string:

H$_2$: W000 S000 T600 N000 A526 M000 S353 A530 U514 Y624 L520 L100 T000 K520 B656 H000 Y000 C500 M230 C614 U150 Y600 H600 T200 N000 [. . .]

The new text H_2 now has length 30316 and its reduced vocabulary ("alphabet") consists of 1625 different "letters". The patterns we consider here are

$\mathcal{W}_3$ = *"to be or not to be"* (T000 B000 O600 N300 T000 B000)
$\mathcal{W}_4$ = *"be it or not"* (B000 I300 O600 N300)
$\mathcal{W}_5$ = *"who is the king"* (W000 I200 T000 K520)

With distances all taken at $d = 100$, the observed number of occurrences and the observed/expected ratios Ω/E are then found to be

$$\mathcal{W}_3: \ \Omega = 15767, \Omega/E = 0.68; \quad \mathcal{W}_4: \ \Omega = 238, \Omega/E = 0.60;$$
$$\mathcal{W}_5: \ \Omega = 1038, \Omega/E = 1.40.$$

Examples like the ones above could be multiplied *ad libitum*. The overall conclusion of such observations is the following. For expected values well above 1, and for gaps that are longer than the short-term correlations of the text, the mean value estimates of the number of occurrences are quite faithful to reality; fluctuations from the mean value do then help discriminate between "signal" and "noise" (e.g., compare the left and right graphics in Figures 8 and 9).

7. CONCLUSIONS

The general probabilistic aspects of the statistics of hidden words can now be regarded as fairly well quantified. In particular, we can return to the question that originally motivated the present study, that of finding reliable thresholds. For instance, if false alarms are to be avoided, the problem is rephrased as one of finding a threshold $\alpha_0 = \alpha_0(\mathcal{W}; n, \beta)$ such that

$$\mathbb{P}_n(\Omega > \alpha_0) \leq \beta,$$

where the data are the pattern Ω, the length n of the text, and a given small β (say $\beta = 10^{-5}$). Based on frequencies of letters and the assumption that a memoryless model is (at least roughly) relevant, one can calculate the mean value and the standard deviation coefficients $\pi(\mathcal{W}), \sigma(\mathcal{W})$ by methods of Section 3.3. The Gaussian limits granted by Theorems 2 and 3 then reduce the problem to solving an approximate system, which in the (fully) constrained case reads

$$\alpha_0 = n\pi(\omega) + x_0\sigma(\mathcal{W})\sqrt{n}, \qquad \beta = \frac{1}{\sqrt{2\pi}}\int_{x_0}^{\infty} e^{-t^2/2}\, dt.$$

This system admits of the approximate solution (for β small):

$$\alpha_0 \approx n\pi(\omega) + \sigma(\mathcal{W})\sqrt{2n\log(1/\beta)}. \tag{53}$$

In practical situations, where the probabilistic data model is unknown and data may be rather irregular, some caution should be exercized in applying formulæ blindly and experimentation with what one observes on real data is likely to be a necessity. The moment, central limit, and large deviation results of this paper at least provide a firm conceptual basis under which one can interpret the facts and should permit a fine tuning of pragmatically developed threshold formulæ stemming from (53).

Concerning open problems, an intriguing question is that of quantifying the speed of convergence to the Gaussian limit as well as large deviations in the (fully or partly) unconstrained cases. The corresponding questions appear to be related to products of random matrices and to the difficult case of random walks on nilpotent Lie groups; see Guivarc'h's paper [22] for context and references. An alternative approach has been very recently proposed by Janson [26].

Finally, as already mentioned, some of the results developed here (the analysis of the first two moments as well as concentration of distribution) have recently been shown to hold [9] for Markovian sources and more generally for all dynamical sources in the sense of Vallée. This points to possible extensions of the present work in the direction of more realistic data models than the memoryless case that has been considered here.

Acknowledgments. We thank M. Atallah (Purdue U.) for introducing us to the intrusion detection problem that motivated this study. We are grateful to Yves Guivarc'h for several discussions relative to an earlier version of the paper, to Svante Janson for his interest in this research, and to Loick Lhote for his careful reading of Sections 2 and 3.

References

[1] S.-S Abhyankar, *Algebraic geometry for scientists and engineers.* American Mathematical Society, 1990.
[2] A. Aczel, *The Mystery of the Aleph. Mathematics, the Kabbalah, and the Search for Infinity*, Four Walls Eight Windows, New York, 2000.
[3] A. Apostolico and M. Atallah, Compact Recognizers of Episode Sequences, Submitted to *Information and Computation.*
[4] E. A. Bender, Central and local limit theorems applied to asymptotic enumeration. *Journal of Combinatorial Theory 15* (1973), 91–111.
[5] E. A. Bender and F. Kochman, The distribution of subword counts is usually normal. *European Journal of Combinatorics 14* (1993), 265–275.
[6] E. A. Bender, L. B. Richmond, and S. G. Williamson, Central and local limit theorems applied to asymptotic enumeration. III. Matrix recursions. *Journal of Combinatorial Theory,* Series A *35*, 3 (1983), 264–278.
[7] P. Billingsley, *Probability and Measure*, Second Edition, John Wiley & Sons, New York, 1986.

[8] L. Boasson, P. Cegielski, I. Guessarian, and Yuri Matiyasevich, Window-Accumulated Subsequence Matching Problem is Linear, In *Proceedings of the Eighteenth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems: PODS 1999*, ACM Press, 327–336, 1999.

[9] J. Bourdon and B. Vallée, Generalized Pattern Matching Statistics. In *Mathematics and computer science* (Colloquium Proceedings, Versailles, 2002), B. Chauvin et al. Editors, Birkhäuser Verlag, 2002, pp. 229–245.

[10] J. Clément, P. Flajolet, and B. Vallée, Dynamical Sources in Information Theory: A General Analysis of Trie Structures, *Algorithmica*, 29, 307–369, 2001.

[11] M. Crochemore and W. Rytter, *Text Algorithms*, Oxford University Press, New York, 1994.

[12] G. Das, R. Fleischer, L. Gasieniec, D. Gunopulos, and J. Kärkkäinen, Episode Matching, In *Combinatorial Pattern Matching, 8th Annual Symposium*, *Lecture Notes in Computer Science* vol. 1264, 12–27, 1997.

[13] F. den Hollander, *Large deviations*. American Mathematical Society, Providence, RI, 2000.

[14] P. Flajolet, Y. Guivarch, W. Szpankowski, and B. Vallée. Hidden Pattern Statistics. In *Automata, Languages, and Programming* (Proceedings of the 28th ICALP Conference), *Lecture Notes in Computer Science*, vol. 2076 (2001), pp. 152–165.

[15] P. Flajolet, P. Kirschenhofer, and R. Tichy, Deviations from uniformity in random strings. *Probability Theory and Related Fields 80* (1988), 139–150.

[16] P. Flajolet, and R. Sedgewick, *Analytic Combinatorics*, In prep., 2002. (Available electronically at `http:algo.inria.fr/flajolet/Publications`.)

[17] P. Flajolet and M. Soria General combinatorial schemas: Gaussian limit distributions and exponential tails. *Discrete Mathematics 114* (1993), 159–180.

[18] F. R. Gantmacher, *Matrizentheorie*. Deutscher Verlag der Wissenschaften, Berlin, 1986. A translation of the Russian original *Teoria Matriz*, *Nauka*, Moscow, 1966.

[19] B. V. Gnedenko and A. N. Kolmogorov, A. N., *Limit Distributions for Sums of Independent Random Variables*. Addison-Wesley, 1968.

[20] L. Guibas and A. M. Odlyzko, Periods in Strings, *J. Combinatorial Theory Ser. A*, 30, 19–43, 1981.

[21] L. Guibas and A. M. Odlyzko, String Overlaps, Pattern Matching, and Nontransitive Games, *J. Combinatorial Theory Ser. A*, 30, 183–208, 1981.

[22] Y. Guivarc'h, Marches aléatoires sur les groupes, *Fascicule de probabilités*, Publ. Inst. Rech. Math. Rennes, 2000.

[23] Hsien-Kuei Hwang, Large deviations for combinatorial distributions: I. Central limit theorems, *The Annals of Applied Probability*, 6, 297–319, 1996.

[24] Hsien-Kuei Hwang, On convergence rates in the central limit theorems for combinatorial structures, *European Journal of Combinatorics*, 19, 329–343, 1998.

[25] S. Janson. *Gaussian Hilbert spaces*, Cambridge University Press, 1997.

[26] S. Janson, Large deviations for sums of partially dependent random variables, preprint 2002.

[27] T. Kato. *Perturbation theory for linear operators*, Springer-Verlag, 1980.

[28] D. E. Knuth, *The Art of Computer Programming, Fundamental Algorithms*, Vol. 1, Third Edition, Addison-Wesley, Reading, MA, 1997.

[29] D. E. Knuth, *The Art of Computer Programming. Sorting and Searching*, Vol. 3, Second Edition, Addison-Wesley, Reading, MA, 1998.

[30] G. Kucherov and M. Rusinowitch, Matching a Set of Strings with Variable Length Don't Cares, *Theoretical Computer Science* 178, 129–154, 1997.

[31] S. Kumar and E.H. Spafford, A Pattern-Matching Model for Intrusion Detection, *Proceedings of the National Computer Security Conference*, 11–21, 1994.

[32] P. Nicodème, B. Salvy, and P. Flajolet, Motif Statistics, *European Symposium on Algorithms*, *Lecture Notes in Computer Science*, No. 1643, 194–211, 1999.

[33] M. Régnier and W. Szpankowski, On the Approximate Pattern Occurrences in a Text, *Proc. Compression and Complexity of SEQUENCE'97*, IEEE Computer Society, 253–264, Positano, 1997.

[34] M. Régnier and W. Szpankowski, On Pattern Frequency Occurrences in a Markovian Sequence, *Algorithmica*, 22, 631-649, 1998.

[35] I. Rigoutsos, A. Floratos, L. Parida, Y. Gao and D. Platt, The Emergence of Pattern Discovery Techniques in Computational Biology, *Metabolic Engineering*, 2, 159-177, 2000.

[36] R. Sedgewick and P. Flajolet, *An Introduction to the Analysis of Algorithms*, Addison-Wesley, Reading, MA, 1995.

[37] J. M. Steele, *Probability Theory and Combinatorial Optimization*, SIAM, Philadelphia, 1997.

[38] W. Szpankowski, *Average Case Analysis of Algorithms on Sequences*, John Wiley & Sons, New York, 2001.

[39] B. Vallée, Dynamical Sources in Information Theory: Fundamental Intervals and Word Prefixes, *Algorithmica*, 29, 262–306, 2001.

[40] A. Vanet, L. Marsan, and M.-F. Sagot, Promoter sequences and algorithmical methods for identifying them, *Res. Microbiol.*, 150, 779-799, 1999.
[41] M. Waterman, *Introduction to Computational Biology*, Chapman and Hall, London, 1995.
[42] A. Wespi, H. Debar, M. Dacier, and M. Nassehi, Fixed vs. Variable-Length Patterns For Detecting Suspicious Process Behavior, *J. Computer Security*, 8, 159-181, 2000.
[43] S. Wu and U. Manber, Fast Text Searching Allowing Errors, *Comm. ACM*, 35:10, 83–991, 1995.