

# Hidden Word Statistics

PHILIPPE FLAJOLET

*INRIA-Rocquencourt, Le Chesnay, France*

WOJCIECH SZPANKOWSKI

*Purdue University, West Lafayette, Indiana*

AND

BRIGITTE VALLÉE

*Université de Caen, Caen Cedex, France, and  
INRIA-Rocquencourt, Le Chesnay, France*

**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 motivations to study this problem come from an attempt at finding a reliable threshold for intrusion detections, from textual data processing applications, and from molecular biology.

Categories and Subject Descriptors: F.2.2. [**Analysis of Algorithms and Problem Complexity**]: Non-numerical Algorithms and Problems—*Computations on discrete structures*; G.2.1 [**Discrete Mathematics**]: Combinatorics—*Generating functions; counting problems*

---

An extended abstract of this article has appeared as FLAJOLET, P., GUIVARCH, Y., SZPANKOWSKI, W., AND VALLÉE, B., Hidden pattern statistics, in *Automata, Languages, and Programming (Proceedings of the 28th ICALP Conference)*. Lecture Notes in Computer Science, vol. 2076. Springer-Verlag, New York, 2001, pp. 152–165.

This research was supported in part by the ALCOM-FT Project (#IST-1999-14186) of the European Union; the NSF Grants CCR-0208709, CCF-0513636, and DMS-0503742; NIH Grant R01 GM068959-01; and AFOSR Grant FA8655-04-1-3074.

Authors’ addresses: P. Flajolet, Algorithms Project, INRIA-Rocquencourt, 78153 Le Chesnay, France, e-mail: Philippe.Flajolet@inria.fr, W. Szpankowski, Department of Computer Science, Purdue University, West Lafayette, IN 47907, e-mail: spa@cs.purdue.edu; B. Vallée, GREYC, Université de Caen, F-14032 Caen Cedex, France, and Algorithms Project, INRIA-Rocquencourt, 78153 Le Chesnay, France, e-mail: Brigitte.Vallee@info.unicaen.fr.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or direct commercial advantage and that copies show this notice on the first page or initial screen of a display along with the full citation. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, to redistribute to lists, or to use any component of this work in other works requires prior specific permission and/or a fee. Permissions may be requested from Publications Dept., ACM, Inc., 1515 Broadway, New York, NY 10036 USA, fax: +1 (212) 869-0481, or [permissions@acm.org](mailto:permissions@acm.org).

© 2006 ACM 0004-5411/06/0100-0147 \$5.00

General Terms: Algorithms

Additional Key Words and Phrases: Pattern matching, words, subsequences, languages, combinatorial calculus, generating functions, complex asymptotics, De Bruijn graph, discrete probability, distributional analysis

## 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 [Crochemore and Rytter 1994] 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 [Bender and Kochman 1993; Flajolet et al. 1988; Guibas and Odlyzko 1981a, 1981b; Lothaire 2005; Nicodème et al. 1999; Régnier and Szpankowski 1997, 1998; Waterman 1995]. For instance, Guibas and Odlyzko [1981a, 1981b] have revealed the fundamental rôle played by autocorrelation vectors and their associated polynomials. Régnier and Szpankowski [1997, 1998] established that the number of occurrences of a string is asymptotically normal under a diversity of models that include Markov chains. Nicodème et al. [1999] showed generally that the number of *places* (positions) in a random text at which a “motif” (i.e., a general regular expression pattern) terminates is asymptotically normally distributed. A survey on various string matching problems and their analyses can be found in Lothaire [2005, Chap. 7].

In sequence comparison, 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  $1 \leq i_1 < i_2 < \dots < i_m \leq n$  such that  $t_{i_1} = w_1, t_{i_2} = w_2, \dots, t_{i_m} = w_m$ . We also say that the word  $w$  is “*hidden*” in the text; thus, we call this the *hidden pattern* problem. For example, `date` 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

$$\text{ab}\#_2\text{r}\#\text{ac}\#\text{a}\#\text{d}\#_4\text{a}\#\text{br}\#\text{a}; \quad (1)$$

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.

1.1. 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 Apostolico and Atallah [2002], Kumar and Spafford [1994], and Wespi et al. [2000]. 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—this is due to the distributed nature of intervening events. 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. By knowing the most likely number of occurrences and the probability of deviating from it, we can determine a plausible threshold such that we are only likely to miss real attacks with a small probability (see the large deviation estimates of Theorem 3 and the remarks at the end of Section 7).

When conducting our experiments on Shakespeare’s works (hardly a Bernoulli text), we found that the counts for (unconstrained) hidden words prove often to be barely distinguishable, up to stochastic noise, from what the Bernoulli model predicts. Typically, for the first two examples discussed in Section 6, the discrepancy between what is observed in the actual text and a Bernoulli model based on actual letter occurrences is about 5%. The fact that some leeway is afforded to spacings between letters appears largely to nullify the effect of correlations arising under various source models. Indeed, there are some recent experimental results conducted on real-life audit files [Atallah et al. 2003; Gwadera et al. 2004] suggesting that our approach is useful even though the source is not memoryless and, in [Gwadera et al. 2005], it was experimentally verified that Markov sources (of higher order) induce only a fairly slight modification of the values of predicted thresholds. The thresholds computed here thus appear to be endowed with a certain amount of robustness, and our analysis under a simple Bernoulli model can certainly serve as the starting point of further studies, either experimental or theoretical.

*Molecular biology* provides another important source of applications [Rigoutsos et al. 2000; Vanet et al. 1999; Waterman 1995]. 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. For example, it is well known that certain regulatory sites in organisms such as *E. coli* and yeast have the form of “doublets”, consisting of two short words separated by a variable-length spacer of arbitrary composition. If the words in the doublet are sufficiently well conserved, our model can adequately describe the statistics of such motifs. Similarly, a collection of motifs in a promoter region with conserved order and orientation might be modeled as a series of spaced words. In

order to overcome a limitation of our analysis for finding biological motifs (which are not perfectly conserved words), one needs to extend our approach to approximate hidden patterns: A preliminary step in this direction was already taken by Bourdon and Vallée [2002], who furthermore consider source models encompassing Markov chains and models with unbounded correlations that arise from dynamical systems theory. Finally, 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 [Steele 1997].

We, computer scientists and mathematicians, are certainly not the first who invented hidden words and hidden meaning [Aczel 2000]. 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!). In fact, recent years have seen a resurgence of interest in finding hidden messages in the Bible and similar texts. See the “Bible Codes” controversy started by the publication of Witztum et al. [1994], with refutations provided by McKay et al. [1999]. (See also Brendan McKay’s informative site, *Scientific Refutation of the Bible Codes*.) To this debate, we only contribute a common-sense remark: our convergence in probability results confirm that any predetermined pattern is likely to be detected in any text as soon as its mean number of occurrences is sufficiently high, which itself is granted as soon as enough leeway is allowed in the choice of spacings between letters, of spelling variants, and so on.

Recent versions of these activities 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 Wu and Manber [1995], Kucherov and Rusinowitch [1997], 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 Boasson et al. [1999], Das et al. [1997], Kucherov and Rusinowitch [1997], and Wu and Manber [1995] 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 quantify *subsequence statistics*, that is, we determine 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 or stops (probably) being meaningful.

1.2. 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, that is, 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  $\Omega$ . We prove in Theorem 1 that the number of occurrences  $\Omega$  in a random text of size  $n$  has expectation and variance given by

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

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 following (incomplete) list of references: [Bender and Kochman 1993; Guibas and Odlyzko 1981a, 1981b; Nicodème et al. 1999; Régnier and Szpankowski 1997, 1998; Waterman 1995]. It is well known that for string matching, the variance coefficient  $\sigma^2$  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).

1.3. 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 [Flajolet and Sedgewick 2005]. 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.

## 2. Framework

We fix an alphabet  $\mathcal{A} := \{a_1, a_2, \dots, a_r\}$ . The set of all possible texts is  $\mathcal{A}^*$  (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*.

**2.1. POSITIONS AND OCCURRENCES.** An  $m$ -tuple  $I = (i_1, i_2, \dots, i_m)$  ( $1 \leq i_1 < i_2 < \cdots < i_m$ ) satisfies the constraint  $\mathcal{D}$  if  $i_{j+1} - i_j \leq d_j$ , 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  $i_m \leq n$ . 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  $I = (i_1, i_2, \dots, i_m)$  of  $\mathcal{P}_n(\mathcal{D})$  and a text  $T = t_1 t_2 \cdots t_n$  for which  $t_{i_1} = w_1, t_{i_2} = w_2, \dots, t_{i_m} = w_m$ . Thus, what we call an occurrence is a text augmented with the distinguished positions at which the pattern occurs. The number  $\Omega$  of occurrences of pattern  $w$  in text  $T$  subject to the constraint  $\mathcal{D}$  is then a sum of characteristic variables

$$\Omega(T) = \sum_{I \in \mathcal{P}_T(\mathcal{D})} X_I(T), \quad \text{with } X_I(T) := \llbracket w \text{ occurs at position } I \text{ in } T \rrbracket. \quad (2)$$

There, Iverson’s bracket convention is used:

$$\llbracket B \rrbracket = \begin{cases} 1 & \text{if the property } B \text{ holds,} \\ 0 & \text{otherwise.} \end{cases} \quad (3)$$

**2.2. 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  $1 \leq b \leq m$ . 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  $\mathcal{W}_1, \mathcal{W}_2, \dots, \mathcal{W}_b$ . 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  $((\mathcal{W}_1, \mathcal{D}_1), (\mathcal{W}_2, \mathcal{D}_2), \dots, (\mathcal{W}_b, \mathcal{D}_b))$ .

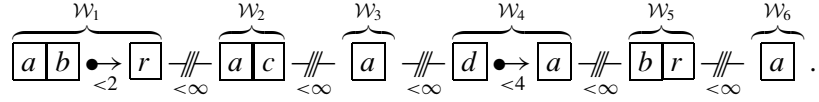
*Example 1.* With the example (1) described in the introduction, namely,

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

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

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

In more figurative terms, this is described as follows (with springs  $\overline{\text{---}}$  representing unbounded gaps,  $\bullet \rightarrow$  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  $I = (i_1, i_2, \dots, i_m)$  of  $\mathcal{W}$  subject to constraint  $\mathcal{D}$  gives rise to  $b$  suboccurrences,  $I^{[1]}, I^{[2]}, \dots, I^{[b]}$ , the  $r$ th term  $I^{[r]}$  representing an occurrence of  $\mathcal{W}_r$  subject to constraint  $\mathcal{D}_r$ . The  $r$ th 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,

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

this is extended to a general hidden pattern specification as

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

**2.3. PROBABILISTIC MODEL.** As regards the probabilistic model, we consider a *memoryless source* that emits symbols of the text independently from the fixed finite alphabet  $\mathcal{A} = \{a_1, a_2, \dots, a_r\}$  and denote by  $p_\alpha$  ( $0 < p_\alpha < 1$ ) 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  $\Omega$  to  $\mathcal{A}^n$ , denoted by  $\Omega_n$  whenever dependency on size needs to be made explicit, becomes a *random variable* defined on  $\mathcal{A}^n$ . Then,  $\Omega_n$ , 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  $\mathcal{A}^*$ , namely  $\Xi$  defined in (11) and  $\tilde{\Xi}$  defined in (24). For any such parameter, say  $U$ , we shall adopt similar notations:  $U$  is the parameter defined on  $\mathcal{A}^*$ ,  $U_n$  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  $\mathcal{A}^n$ .

2.4. GENERATING FUNCTIONS. Throughout this article, we shall consider structures superimposed on words. For a class  $\mathcal{C}$  of structures and given a weight function  $\gamma$  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  $\omega$ . In particular, the usual counting generating function corresponds to the constant weight  $\gamma \equiv 1$ . Then,<sup>1</sup>  $C_n = [z^n]C(z)$  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 unions, and Cartesian products. Thus, a minimal set of rules must first be given in order to translate such basic constructions; see Flajolet and Sedgewick [2005], Sedgewick and Flajolet [1995], and Szpankowski [2001] for a general framework.

Take  $\mathcal{A}, \mathcal{B}, \mathcal{C}$  to be weighted sets with respective weights  $\alpha, \beta, \gamma$ . 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  $\gamma$  on  $\mathcal{C}$  is inherited from the weights  $\alpha, \beta$  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:

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

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

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



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

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

since one has

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

(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  $\mathcal{A}^*$  is

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

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,

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

### 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  $I = (i_1, i_2, \dots, i_m)$  of  $\mathcal{P}_n(\mathcal{D})$  and a text  $T = t_1 t_2 \dots t_n$  for which  $t_{i_1} = w_1, t_{i_2} = w_2, \dots, t_{i_m} = w_m$ . 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

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

and, with the definition of  $\Omega$ ,

$$O(z) = \sum_T \Omega(T) \pi(T) z^{|T|} = \sum_n \mathbb{E}_n[\Omega] z^n. \quad (7)$$

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

$$\mathcal{O} = \mathcal{A}^* \times \{w_1\} \times \mathcal{A}^{<d_1} \times \{w_2\} \times \mathcal{A}^{<d_2} \times \dots \times \{w_{m-1}\} \times \mathcal{A}^{<d_{m-1}} \times \{w_m\} \times \mathcal{A}^*. \quad (8)$$

There, for  $d < \infty$ ,  $\mathcal{A}^{<d}$  denotes the collection of all words of length strictly less  $d$ , that is,  $\mathcal{A}^{<d} := \bigcup_{i < d} \mathcal{A}^i$ , whereas, for  $d = \infty$ ,  $\mathcal{A}^{<\infty}$  denotes the collection of all finite words, that is,  $\mathcal{A}^{<\infty} := \mathcal{A}^* = \bigcup_{i < \infty} \mathcal{A}^i$ . 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

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

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

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

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

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

and a complete asymptotic expansion could be easily obtained. This symbolic derivation of mean values extends the case of standard string matching exposed in Sedgewick and Flajolet [1995, 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  $\Xi$  defined, for each  $n$ , as

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

The second moment of the centered variable  $\Xi$  equals the variance of  $\Omega$  and with the centered variables defined above by (11), one has

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

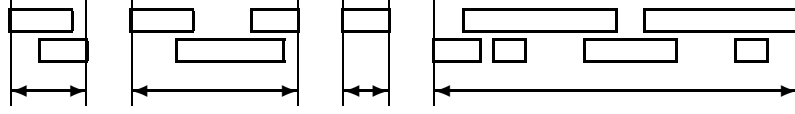


FIG. 1. A pair of occurrences  $I, J$  with  $b = 6$  blocks each and the joint aggregates; the number of degrees of freedom is here  $\beta(I, J) = 4$ .

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 Eq. (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  $(I, J)$ , depending on whether 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  $\ell$  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  $e(I, J)$

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

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

**3.2.1. Aggregates and Degrees of Freedom of Pairs of Positions.** As it turns out in the analysis, asymptotic behavior 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.

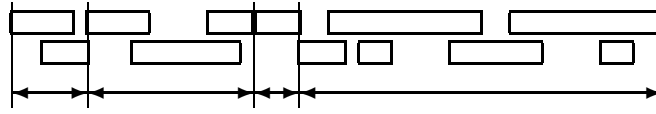


FIG. 2. A full pair of occurrences  $I, J$  with  $b = 6$  blocks each.

*Example 2.* Consider the pattern  $\mathcal{W} = \boxed{a\#_3b\#_4r} \# \boxed{a\#_4c}$  composed of two blocks. The text `aarbarbccaaracc` 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  $\mathcal{O}_2$  in accordance with the value of  $\beta(I, J)$  and write

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

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

$$\mathcal{O}_2(z) = \mathcal{O}_2^{[1]}(z) + \mathcal{O}_2^{[2]}(z) + \mathcal{O}_2^{[3]}(z) + \dots$$

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

**3.2.2. 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  $2\ell$ , where  $\ell$  is the length of the constraint  $\mathcal{D}$ .)

*Example 3.* Consider the pattern  $\mathcal{W} = a\#_3b\#_4r\#a\#_4c$ . The text `aarbarbccaaracc` 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 corresponding aggregates are  $\alpha(I') = \{[1, 6], [12, 13]\}$ ,  $\alpha(J') = \{[5, 11], [12, 14]\}$  so that  $\alpha(I', J') = \{[1, 11], [12, 14]\}$ . This is an example of a full pair of occurrences with 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)

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

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

$$\mathcal{O}_2^{[p]}(z) = \left(\frac{1}{1-z}\right)^{2b-p+1} \times \mathcal{B}_2^{[p]}(z).$$

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 Guibas and Odlyzko [1981a, 1981b]. 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  $\Omega$  of a pattern  $\mathcal{W}$  subject to constraint  $\mathcal{D}$  satisfy*

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

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

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

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  $1 \leq r, s \leq b$  for which the  $r$ th block  $B^{[r]}$  of  $\alpha(I)$  and the  $s$ th block  $C^{[s]}$  of  $\alpha(J)$  intersect.

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

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

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

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

Bourdon and Vallée [2002] 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 Bourdon and Vallée [2002] to hold for a wide class of sources encompassing memoryless and Markov sources—the *dynamical sources* in the sense of Vallée [Clément et al. 2001; Vallée 2001].

3.3. GENERALIZED AUTOCORRELATIONS AND VARIANCES. In this section, we re-examine the variance coefficient, for which formulas have been provided earlier; see Eq. (14) in 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 Guibas and Odlyzko [1981a, 1981b] and Szpankowski [2001]).

3.3.1. *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

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

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  $1 \leq r, s \leq b$  for which the  $r$ th block  $B^{[r]}$  of  $\alpha(I)$  and the  $s$ th 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  $j$ th block); (iii) finally the relative positions of the blocks  $B^{[r]}$  and  $C^{[s]}$ .

3.3.2. *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

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

3.3.3. *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 pattern  $\mathcal{W}$ .) 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  $\mathcal{C} = (c_1, c_2, \dots, c_{i-1})$ ; in the same vein, the 1-block pattern  $\mathcal{S}$  has  $j$  symbols, so that constraint  $\mathcal{D}$  is of the form  $\mathcal{D} = (d_1, d_2, \dots, d_{j-1})$ .

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

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

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

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

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

$$(19) \quad \begin{aligned} \mathcal{C}(t, k) &:= \{u; 1 \leq u \leq t-1, u \geq t - c_{k-1}\} \\ \mathcal{D}(v, \ell) &:= \{w; 1 \leq w \leq v-1, w \geq v - d_{\ell-1}\}. \end{aligned}$$

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

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

The initialization conditions are

$$(22) \quad \begin{aligned} Y[1, 0, 1, 0] &:= 1; & Y[0, 1, 0, 1] &:= 1; & Y[1, 1, 1, 1] &:= \llbracket r_1 = s_1 \rrbracket; \\ X[1, 0, 1, 0] &:= p(r_1); & Y[0, 1, 0, 1] &:= p(s_1); & X[1, 1, 1, 1] &:= p(r_1) \llbracket r_1 = s_1 \rrbracket, \end{aligned}$$

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

$$(23) \quad X[t, v, 1, 0] = Y[t, v, 1, 0] = X[t, v, 0, 1] = X[t, v, 0, 1] = Y[t, v, 1, 1] = X[t, v, 1, 1] = 0.$$

FIG. 3. The formulas summarizing the computation of the variance coefficient.

the set  $\mathcal{B}$  is equal to

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

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

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

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 formulas of this section are grouped inside Figure 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  $j_\ell$  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  $s_\ell$  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  $\ell$  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  $\kappa$ , while the second one is the total weight of this set and is used for computing the first term  $A$  of  $\kappa$ ; see Eq. (18) in Figure 3.

The fundamental formulas 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 Figure 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 Figure 3, Eq. (23).

3.3.4. *The Case of a String and the Relation with Autocorrelation Polynomials.* Here  $\mathcal{W} = w_1 w_2 \cdots w_m$  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  $1 \leq i \leq j \leq m$ , we denote by  $\mathcal{W}[i, j]$  the substring  $w_i w_{i+1} \cdots w_j$ . The autocorrelation set  $K_{\mathcal{W}} \subset [1 \cdot m]$  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  $\mathcal{W}[1 \cdot k] = \mathcal{W}_{I \cap J}$ .

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

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

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  $\Omega_n$  appropriately centered and scaled tends to the normal distribution. We consider the following standardized random variable  $\tilde{\Xi}$  which is defined for each  $n$  by

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

where  $b$  is the number of blocks of the constraint  $\mathcal{D}$ . We shall show that  $\tilde{\Xi}$  behaves asymptotically as a normal variable with mean 0 and standard deviation  $\sigma$ . By the classical *moment convergence theorem* [Billingsley 1986, Theorem 30.2] this is established once all moments of  $\tilde{\Xi}_n$  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  $s \geq 0$

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

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

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

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

**THEOREM 2.** *The random variable  $\Omega$  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*

$$\lim_{n \rightarrow \infty} \mathbb{P}_n \left\{ \frac{\Omega - \mathbb{E}_n[\Omega]}{\sqrt{\mathbb{V}_n[\Omega]}} \leq x \right\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x \exp(-t^2/2) dt \quad (27)$$

**PROOF.** The proof below is combinatorial; it basically reduces to grouping and enumerating adequately the various combinations of indices in the sum that expresses  $\mathbb{E}_n[\Xi^r]$  see Figure 4. 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  $\Xi^r$  yields

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

An  $r$ -tuple of sets  $(I_1, \dots, I_r)$  in  $\mathcal{P}^r(\mathcal{D})$  is said to be *friendly* if each  $I_k$  intersects at least one other  $I_\ell$ , 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  $(I_1, \dots, I_r)$  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

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

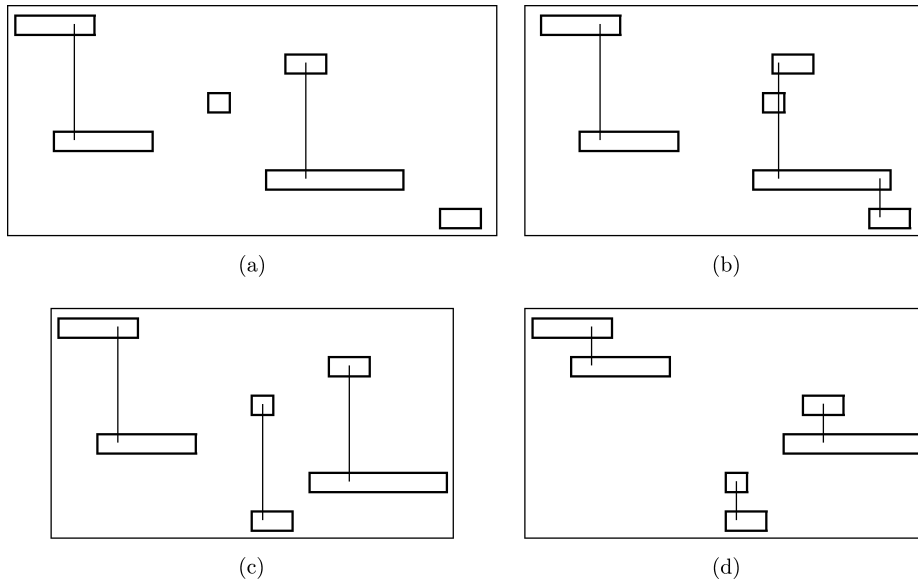


FIG. 4. Various types of families of occurrence positions for  $r = 2s = 6$ : (a) an unfriendly family in  $\mathcal{P}^r$ ; (b) a friendly family in  $\mathcal{Q}^{(r)}$ ; (c) a friendly family with maximal number of joint blocks in  $\mathcal{Q}_\star^{(2s)}$ ; (d) a friendly family with maximal number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

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.

4.1. 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  $\alpha(I_1) \cup \dots \cup \alpha(I_r)$ . 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

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

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

4.2. 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  $I_\ell$ . Furthermore, if the blocks of  $\alpha(I_h)$  are denoted by  $B_h^{[u]}$ ,  $1 \leq u \leq b$ , 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  $\tau$  such that  $\tau(k) = \ell$  and  $\tau(\ell) = k$  for all pairs of indices  $(\ell, k)$  for which  $I_k$  and  $I_\ell$  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  $\tau$  is the standard one with cycles  $(1, 2), (3, 4), \dots$ ; 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. Flajolet and Sedgewick [2005]), the equality

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

entails that we can work now solely with standard families.

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

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

In view of the formulas (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.  $\square$

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

of Billingsley [1986]. Proceeding along different tracks, Janson [2004] has related this particular case to his treatment of  $U$ -statistics via Gaussian Hilbert spaces; see Chapter XI of Janson [1997] 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  $\mathcal{D} = (d_1, d_2, \dots, d_{m-1})$ , 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  $\mathcal{A} = \{a_1, \dots, a_r\}$ , and, in order to avoid trivialities, we assume its cardinality to be at least 2. Also,  $\Omega$  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  $\Omega$  to denote the corresponding random variable over the probability space  $\mathcal{A}^n$  equipped with the Bernoulli (memoryless) model.

The mean and variance of  $\Omega$  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  $\Omega$  satisfies a Central Limit Law with speed of convergence  $1/\sqrt{n}$ :*

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

- (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*

$$\begin{cases} \lim_{n \rightarrow \infty} \frac{1}{n} \log \mathbb{P}_n \left( \frac{\Omega}{n} \leq x \right) = -I(x) & \text{if } 0 < x < D\pi(\mathcal{W}) \\ \lim_{n \rightarrow \infty} \frac{1}{n} \log \mathbb{P}_n \left( \frac{\Omega}{n} \geq x \right) = -I(x) & \text{if } D\pi(\mathcal{W}) < x < \eta \end{cases}, \quad (33)$$

*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 Eq. (49) below).*

$$I(x) = -\log \frac{\lambda(\zeta)}{\zeta^x} \quad \text{with } \zeta \equiv \zeta(x) \text{ defined by } \frac{\zeta \lambda'(\zeta)}{\lambda(\zeta)} = x, \quad (34)$$

*where  $\lambda(u)$  is the largest eigenvalue of  $T(u)$ .*

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

$$\sup_k \left| \mathbb{P}_n(\Omega = k) - \frac{1}{\sigma(\mathcal{W})\sqrt{n}} \frac{\exp(x(k)^2/2)}{\sqrt{2\pi}} \right| = o\left(\frac{1}{\sqrt{n}}\right), \quad x(k) = \frac{k - D\pi(\mathcal{W})n}{\sigma(\mathcal{W})\sqrt{n}}. \quad (35)$$

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 (35).

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 [Bender et al. 1983]; (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 1973 and Hwang [1996, 1998]) for the probability generating function of  $\Omega$  is then inferred, see Eq. (46). 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 [Bender 1973; Flajolet and Soria 1993; Hwang 1996, 1998]. 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 (35).

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 et al. [1983]. 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 Flajolet et al. [1988]. Bender and Kochman [1993] make an implicit use of this construction combined with the central and local limit theorems of Bender et al. [1983] 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 Bender and Kochman [1993] 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 [Bender and Kochman 1993; Bender et al. 1983; Flajolet et al. 1988; Hwang 1996, 1998], 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  $\Delta$  the quantity  $r^\delta$ . There exist a matrix  $T(u)$  of dimension  $\Delta \times \Delta$  and two column vectors  $x(u)$ ,  $y$  of dimension  $\Delta$  such that*

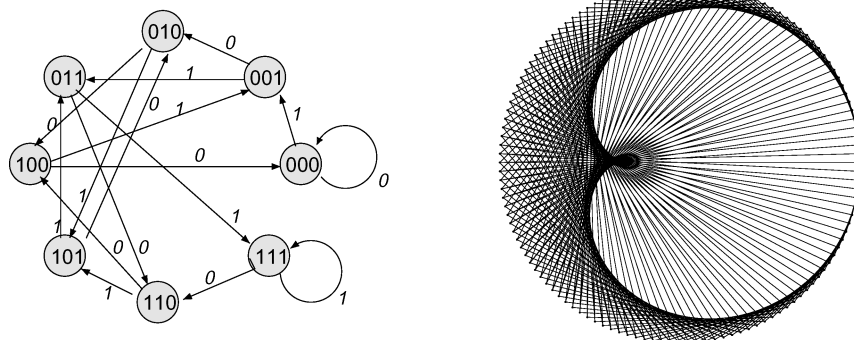


FIG. 5. The de Bruijn graph corresponding to the binary alphabet  $\mathcal{A} = \{0, 1\}$  and to block size equal to  $\ell$  has  $2^\ell$  vertices (associated to blocks of length  $\ell$ ) and  $2^{\ell+1}$  edges: the cases displayed are  $\ell = 3$  (left) and  $\ell = 7$  (right).

the probability generating function of the number of occurrences,  $\Omega$  satisfies, for  $n \geq \delta$ ,

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

The entries of  $T(u)$  and  $\mathbf{x}(u)$  are polynomials in  $u$  with nonnegative coefficients. The vector  $\mathbf{y} = (1, \dots, 1)^t$  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  $\delta$  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  $\alpha$  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 with  $\alpha$ ). 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  $\delta$  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  $\Omega$  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  $\alpha$ , that is, all the occurrences of the pattern that end at the letter  $\alpha$ . 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  $\alpha$  is determined by either the pair  $(b, \alpha)$  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  $\alpha$  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 ( $\llbracket \cdot \rrbracket$  is Iverson's bracket):

$$(T(u))_{b,c} := p_\alpha u^{\phi(b,\alpha)} \llbracket b\alpha \in \mathcal{A}c \rrbracket = p_\alpha u^{\Omega(b\alpha) - \Omega(b)} \llbracket b\alpha \in \mathcal{A}c \rrbracket. \quad (36)$$

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  $c$ ,

$$(T(u)^k)_{b,c} = \sum_{f \in \mathcal{A}^k \mid bf \in \mathcal{A}^k c} \pi(f) u^{\Omega(bf) - \Omega(b)}. \quad (37)$$

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

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

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

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:

$$T(u) = \begin{matrix} & \begin{matrix} aa & ab & ba & bb \end{matrix} \\ \begin{matrix} aa \\ ab \\ ba \\ bb \end{matrix} & \begin{pmatrix} p_a & p_b u^2 & 0 & 0 \\ 0 & 0 & p_a & p_b u \\ p_a & p_b u & 0 & 0 \\ 0 & 0 & p_a & p_b \end{pmatrix} \end{matrix}. \quad (38)$$

5.2. SPECTRAL PROPERTIES OF TRANSFER MATRICES. Perron–Frobenius theory together with its analytically 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”  $\psi$  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  $j = 1 \dots \Delta$  be a numbering of the eigenvalues of  $T(u)$  taken so that  $|\lambda_1(u)| \geq |\lambda_2(u)| \geq \dots \geq |\lambda_\Delta(u)|$ . 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

$$\rho(u) = \lim_{n \rightarrow \infty} \|T(u)^n\|^{1/n}. \quad (39)$$

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  $a(u)$  whose entries are all strictly

positive. There exists a complex neighborhood of the real positive axis on which the mappings  $u \rightarrow \lambda(u)$ ,  $u \rightarrow \mathbf{a}(u)$  are well defined and analytic; in addition, all the entries of  $\mathbf{a}(u)$  are nonzero.

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

$$\lambda(1) = 1, \quad \lambda'(1) = D\pi(\mathcal{W}), \quad \lambda''(1) + \lambda'(1) - \lambda'(1)^2 = \sigma^2(\mathcal{W}). \quad (40)$$

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

$$\lim_{u \rightarrow 0^+} \frac{u\lambda'(u)}{\lambda(u)} = 0, \quad \lim_{u \rightarrow +\infty} \frac{u\lambda'(u)}{\lambda(u)} = \eta := \max \left\{ \frac{\psi(\mathcal{C})}{|\mathcal{C}|}; \quad \mathcal{C} \text{ a cycle} \right\}. \quad (41)$$

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

Formula (41) 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  $+\infty$ ). Formulas (40) and (41) are best understood when expressed in terms of the function  $\Lambda(s)$ , which has the character of a cumulant generating function:

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

(The Quasi-Powers approximation of (46) 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  $L$ th 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 Eq. (37). Then, the classical Perron–Frobenius theory of nonnegative matrices applies to matrix  $T(u)$  (see, e.g., Gantmacher [1986, Chap. 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

$$\lambda(u) := \lambda_1(u) > |\lambda_2(u)| \geq |\lambda_3(u)| \geq \dots, \quad (42)$$

as well as  $\rho(u) = \lambda(u)$  for positive  $u$ . Also, by this theory, the eigenvector  $\mathbf{a}(u)$  corresponding to  $\lambda(u)$  has all its components that are strictly positive. Then, by classical (analytic) perturbation theory [Kato 1980, Chap. II], there exists a neighborhood of the real positive axis where the functions  $u \rightarrow \lambda(u)$ ,  $u \rightarrow \mathbf{a}(u)$  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, (46) below, show that the mean and variance of  $\Omega_n$  are related to the first two derivatives of  $\lambda(u)$  at 1. This establishes the relations (40).



We next prove the relation (41). We shall only do so for the maximum, describing the behavior of  $\lambda(u)$  as  $u \rightarrow \infty$ , since the dual relation at 0 follows from similar arguments (based on the minimum of the  $\psi$  values along cycles). Let  $p_n(u)$  denote  $\mathbb{E}_n(u^\Omega)$ . The maximum relation in (41) is equivalent to asserting the coincidence of the combinatorially defined ‘‘cyclic index’’  $\eta$  of the flow  $\psi$  with an analytically defined ‘‘Puiseux index’’  $\varpi$ , 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

$$\eta = \sup_{c \text{ path}} \frac{\psi(p)}{|p|} = \sup_{c \text{ cycle}} \frac{\psi(c)}{|c|}. \tag{43}$$

In other words, the cyclic index determines the worst-case behavior of  $\psi$  on long paths.

The Puiseux index is defined as

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

its existence being guaranteed by general properties of algebraic functions (the Newton–Puiseux theorem). The coincidence of  $\varpi$  and  $\eta$  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

$$u^{-C_3} A^n u^{\eta n} < p_n(u) < u^{C_2} B^n u^{\eta n}, \tag{44}$$

where  $C_2, C_3$  are (unessential) constants. Indeed, taking  $n$ th roots in (44) 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 (44). For the upper bound, observe that the  $\psi$ -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  $\Delta^{n+1}$  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  $\bar{P}^n$ , with  $\bar{P}$  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 (39).)

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

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

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 (46) below. Taking  $n$ th 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  $\psi(C)$  be the total weight of cycle  $C$  in the de Bruijn graph relative to the pattern  $(\mathcal{W}, \mathcal{D})$ . The quantity  $\psi_{(\mathcal{W}, \mathcal{D})} := \gcd\{\psi(C); C \text{ a cycle}\}$  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  $\rho(r \exp i\theta) \leq \rho(r)$ .
- (v) Let  $d = \psi_{(\mathcal{W}, \mathcal{D})}$  be the period of pattern  $(\mathcal{W}, \mathcal{D})$ .
  - (v.a) When  $d = 1$ , then  $\rho(r \exp i\theta) < \rho(r)$  for all  $\theta \in ]0, 2\pi[$ .
  - (v.b) When  $d > 1$ , then  $\rho(r \exp i\theta) = \rho(r)$  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  $j \rightarrow i$ .

(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  ${}^tT(r)$  and  ${}^tT(ru)$ . With (i), there exist a dominant eigenvalue  $\lambda := \lambda(r)$  strictly positive and a dominant eigenvector  $a := a(r)$  of  ${}^tT(r)$  relative to  $\lambda(r)$  whose all entries  $a_j$  are strictly positive. Consider an eigenvalue  $\mu$  of  ${}^tT(ru)$  and an eigenvector  $c$  relative to  $\mu$ . Denote by  $v_j$  the ratio  $c_j/a_j$ . One can always choose vectors  $a$  and  $c$  such that  $\max_{1 \leq j \leq \Delta} |v_j| = 1$ . Suppose that this maximum is attained for some index  $i$ . One has

$$|\mu c_i| = \left| \sum_j p_{i|j}(ru)^{\psi(j,i)} c_j \right| \leq \sum_j p_{i|j} r^{\psi(j,i)} a_j = \lambda a_i, \tag{45}$$

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

(v) Suppose now that the equality  $|\mu| = \lambda$  holds. Then, the previous inequalities (45) all become equalities. First, for all indices  $\ell$  such that  $p_{i|\ell} \neq 0$ , we deduce that  $|c_\ell| = a_\ell$ , so that  $v_\ell$  has modulus 1. For these indices  $\ell$ , we have the same equalities in (45) 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,

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

so that,

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

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 labeled by  $\beta$  that starts at  $\beta^\delta$  is convenient. This proves that  $\mu = \lambda$  and that  $u^{\psi(C)} = 1$  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  $v^d$ .  $\square$

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  $w_1, w_2, \dots, w_{m-1}$  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 nonzero coefficient  $c_i$ . If such an index, does not exist, let  $i := m$ . Then, the cycle that starts at the state  $w_1 w_2 \cdots w_{m-1}$  and whose successive edges are labeled by the  $i$  symbols  $w_{m-i}, w_{m-i+1}, \dots, w_{m-1}$  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  $(\mathcal{W}, \mathcal{D})$  that 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 labeled 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  $w_1, w_2, \dots, w_{m-1}$  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 labeled by successive symbols of  $b$ . Clearly this cycle has a weight equal to 1.  $\square$

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 neighbourhood 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  $\mathbb{E}_n[u^\Omega] = x(u)^t T(u)^{n-\delta} y$ , the estimate

$$\mathbb{E}_n[u^\Omega] = c(u)\lambda(u)^{n-\delta}(1 + O(A^n)), \tag{46}$$

for a nonzero analytic function  $c(u)$ . A uniform approximation like (46) 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 [Bender 1973; Hwang 1996, 1998], as we see now.

5.3.1. *Central Limit Law.* Given a Quasi-Powers approximation valid when  $u$  lies in a complex neighborhood of 1, the classical proof of the central limit theorem for sums of independent random variables [Gnedenko and Kolmogorov 1968] can be mimicked and convergence to the Gaussian distribution results, following Bender [1973] and Hwang [1998]. The speed of convergence is found to be  $O(1/\sqrt{n})$  as results from the Berry–Esseen inequalities; see Hwang [1998] for the general argument. In this way, Eq. (32) in Theorem 3 is established.

5.3.2. *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

$$\sum_{k \leq xn} p_{n,k} = [u^{\lfloor xn \rfloor}] \frac{p_n(u)}{1-u} \leq \frac{p_n(\theta)}{(1-\theta)\theta^{\lfloor xn \rfloor}}, \tag{47}$$

for any fixed  $\theta \in (0, 1)$ . Then, the Quasi-Powers approximation (46) applied to (47) yields

$$\sum_{k \leq xn} p_{n,k} = O\left(\frac{\lambda(\theta)^n}{\theta^{xn}}\right). \tag{48}$$

The next move consists in adopting in (48) the particular value of  $\theta$  that produces the best upper bound in (48). To this effect, define

$$I(x) = -\log \frac{\lambda(\zeta)}{\zeta^x} \quad \text{with } \zeta \equiv \zeta(x) \in (0, 1) \text{ defined by } \frac{\zeta \lambda'(\zeta)}{\lambda(\zeta)} = x. \quad (49)$$

(The existence of  $\zeta$  is guaranteed by (ii) and (iii) of Lemma 2.) Then, the upper bound (48) becomes

$$\frac{1}{n} \log \mathbb{P}_n \left( \frac{\Omega}{n} \leq x \right) \leq -I(x) + o(1). \quad (50)$$

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  $Y_\zeta$  of  $\Omega$  defined by

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

for the particular  $\zeta$  of (49). The shifted  $Y_\zeta$  satisfies a Quasi-Powers approximation in the central region  $u \approx 1$  and is thus asymptotically normal provided

$$\lim_{n \rightarrow \infty} \frac{1}{n} \mathbb{V}(Y_\theta) = \lambda''(\zeta) + \lambda'(\zeta) - \lambda'(\zeta)^2 \quad (51)$$

is nonzero. The limit quantity in (51) represents an analytic function of  $\zeta$  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

$$\frac{1}{p_n(\zeta)} \sum_{xn - \sqrt{n} < j \leq xn} p_{n,k} \zeta^j \quad (52)$$

then tends to a nonzero limit (expressible as a Gaussian error function). Since the weights in (52) are all of the form  $\zeta^{xn} \exp(O(\sqrt{n}))$ , a lower bound on the  $p_{n,k}$  follows. Thanks to this, the inequality in (50) can then be changed to equality, which is what Theorem 3 asserts.

A mirror argument (with  $\zeta$  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  $\zeta$  over the *complete range* of the distribution of  $\Omega_n$ .

**5.3.3. 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  $\Omega$  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 behavior that is excluded in the case when  $T(u)$  stems from a primitive pattern.

Granted Lemma 3, one can estimate the probability distribution of  $\Omega$  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 [Gnedenko and Kolmogorov 1968]. One starts from Cauchy's coefficient integral,

$$\mathbb{P}_n(\Omega = k) = \frac{1}{2i\pi} \int_{|z|=1} p_n(u) \frac{du}{u^{k+1}}, \quad (53)$$

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. (35) 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 Guivarc'h [2000] for a survey of some of the relevant methods. Also, Janson [2004] 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 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, for example, [Kucherov and Rusinowitch 1997] 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

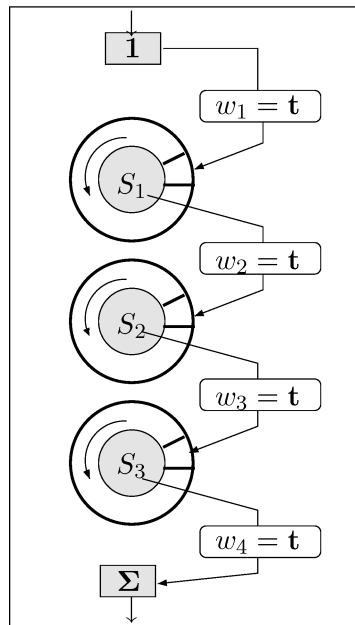


FIG. 6. A hidden-pattern counting machine.

The machine corresponding to a single block ( $b = 1$ ) is composed of wheels and of an output counter  $\Sigma$  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  $w_1, \dots, w_j$  at distance  $0, 1, \dots, (d_j - 1)$  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  $t$  is read from the text, the active sectors are updated: if  $t \neq w_j$ , then the active sector of wheel  $j$  is set to 0; if  $t = w_j$ , 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 1's. The bottom counter  $\Sigma$  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.

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:

$H_0$  : 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:

$H_1$  : whostherenayanswermestandandunfolyourselflonglive [ ... ]

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,

$$\mathcal{W}_0 = \text{thelawisgaussian}, \quad \tilde{\mathcal{W}}_0 = \text{naissuagsiwaleht},$$

$d$	Expected ( $E$ )	$w = \text{thelawisgaussian}$		$\tilde{w} = \text{naissuagsiwalett}$	
		Occurred ( $\Omega$ )	$\Omega/E$	Occurred ( $\Omega$ )	$\Omega/E$
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
$\infty$	1.330E+48	1.36554E+48	1.03	1.38807E+48	1.04

FIG. 7. Observed occurrences ( $\Omega$ ) versus predicted values (expectations,  $E$ ) in the alphabetical characters of Hamlet.

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  $\tilde{\mathcal{W}}_0$  about  $5 \cdot 10^{87}$ . The observed counts, which are  $1.365 \cdot 10^{48}$  and  $1.388 \cdot 10^{48}$  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  $1.330 \cdot 10^{48}$ , 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  $\tilde{w}$  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  $\tilde{\mathcal{W}}$ , 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 Bourdon and Vallée [2002] for such an extension.) Figure 8 further illustrates this by displaying the evolution of the ratios Observed/Expected ( $\Omega/E$ ) as letters in the text are scanned one by one.

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

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

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,



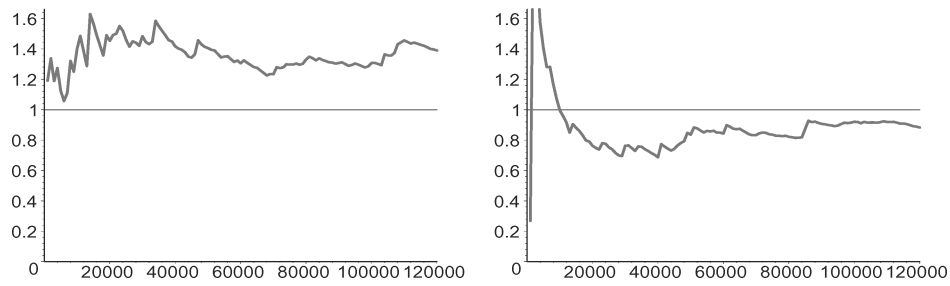


FIG. 8. The evolutions of  $\Omega/E$  for the 1 block patterns  $\mathcal{W}$  (left) and  $\tilde{\mathcal{W}}$  (right) when the separation distance is  $d = 50$ .

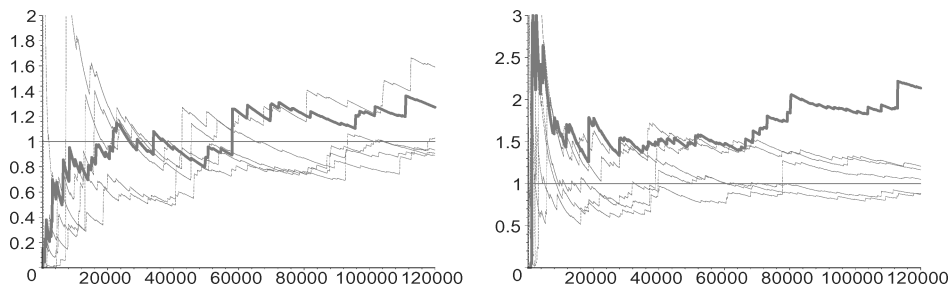


FIG. 9. The evolutions of  $\Omega/E$  for patterns  $\mathcal{W}_1$  (left) and  $\mathcal{W}_2$  (right): a comparison between the original text (thick lines) and 5 randomly permuted versions (dashed lines).

but now relative to five 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?) of words that matter, we simplify the text of Hamlet by applying the Soundex algorithm. (The Soundex algorithm as described by Knuth [1997] 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<sub>2</sub>** :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 [ . . . ]

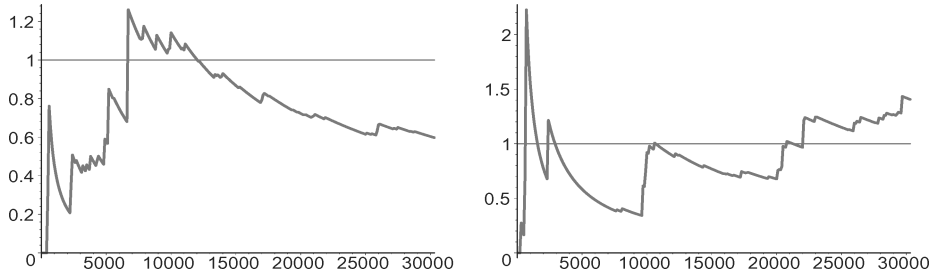


FIG. 10. The Soundex-ed version of Hamlet. The patterns are  $\mathcal{W}_3 \equiv$  “to be or not to be” and  $\mathcal{W}_5 \equiv$  “who is the king” taken with distance  $d = 100$ .

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:

$$\begin{aligned}\mathcal{W}_3 &= \text{“to be or not to be”} \quad (\text{T000 B000 O600 N300 T000 B000}) \\ \mathcal{W}_4 &= \text{“be it or not”} \quad (\text{B000 I300 O600 N300}) \\ \mathcal{W}_5 &= \text{“who is the king”} \quad (\text{W000 I200 T000 K520}).\end{aligned}$$

With distances all taken at  $d = 100$ , the observed number of occurrences and the observed/expected ratios  $\Omega/E$  are then found to be (see also Figure 10):

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

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, at least under the simple model of Bernoulli texts. In particular, we can return to the question that originally motivated the present study, that of finding reliable thresholds. For instance, if false alarms in intrusion detection 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  $\Omega$ , the length  $n$  of the text, and a given small  $\beta$  (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

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

This system admits of the approximate solution (for  $\beta$  small):

$$\alpha_0 \approx n\pi(\omega) + \sigma(\mathcal{W})\sqrt{2n \log(1/\beta)}. \quad (54)$$

In practical situations, where the probabilistic data model is unknown and data may be rather irregular, some caution should be exercised in applying formulas 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 strategies stemming from (54).

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 [2000] for context and references. An alternative approach has been recently proposed by Janson [2004], whose methods provide useful bounds for the probability of deviating from the mean value.

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 [Bourdon and Vallée 2002] 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

- ABHYANKAR, S.-S. 1990. *Algebraic Geometry for Scientists and Engineers*. American Mathematical Society Providence, R.I.
- ACZEL, A. 2000. *The Mystery of the Aleph. Mathematics, the Kabbalah, and the Search for Infinity*, Four Walls Eight Windows, New York.
- APOSTOLICO, A., AND ATALLAH, M. 2002. Compact recognizers of episode sequences. *Inf. Comput.* 174, 2(May), 180–192.
- ATALLAH, M., GWADERA, R., AND SZPANKOWSKI, W. 2003. Reliable detection of episodes in event sequences. In *Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM-03)* (Melbourne, FL), Nov. IEEE Computer Science Press, Los Alamitos, CA, 67–74.
- BENDER, E. A. 1973. Central and local limit theorems applied to asymptotic enumeration. *J. Combinat. Theory* 15, 91–111.
- BENDER, E. A., AND KOCHMAN, F. 1993. The distribution of subword counts is usually normal. *Europ. J. Combinat.* 14, 265–275.
- BENDER, E. A., RICHMOND, L. B., AND WILLIAMSON, S. G. 1983. Central and local limit theorems applied to asymptotic enumeration. III. Matrix recursions. *J. Combinat. Theory, Ser. A* 35, 3, 264–278.
- BILLINGSLEY, P. 1986. *Probability and Measure*, 2nd Ed., John Wiley, New York.
- BOASSON, L., CEGIELSKI, P., GUESSARIAN, I., AND MATIYASEVICH, Y. 1999. Window-accumulated subsequence matching problem is linear. In *Proceedings of the 18th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems PODS*, ACM, New York, 327–336.

- BOURDON, J., AND VALLÉE, B. 2002. Generalized pattern matching statistics. In *Proceedings of the Colloquium on Mathematics and Computer Science* (Versailles, France). B. Chauvin et al. Eds. Birkhäuser Verlag, pp. 229–245.
- CLÉMENT, J., FLAJOLET, P., AND VALLÉE, B. 2001. Dynamical sources in information theory: A general analysis of trie structures. *Algorithmica*, 29, 307–369.
- CROCHEMORE, M., AND RYTTER, W. 1994. *Text Algorithms*. Oxford University Press, New York.
- DAS, G., FLEISCHER, R., GASIENIEC, L., GUNOPULOS, D., AND KÄRKKÄINEN, J. 1997. Episode matching. In *Combinatorial Pattern Matching, 8th Annual Symposium, Lecture Notes in Computer Science*. 12–27.
- DEN HOLLANDER, F. 2000. *Large Deviations*. American Mathematical Society, Providence, RI.
- FLAJOLET, P., KIRSCHENHOFER, P., AND TICHY, R. 1988. Deviations from uniformity in random strings. *Prob. Theory Rel. Fields* 80, 139–150.
- FLAJOLET, P., AND SEDGEWICK, R. 2005. *Analytic Combinatorics*, In prep., (Available electronically at <http://algo.inria.fr/flajolet/Publications/>.)
- FLAJOLET, P., AND SORIA, M. 1993. General combinatorial schemas: Gaussian limit distributions and exponential tails. *Discr. Math.* 114, 159–180.
- GANTMACHER, F. R. 1986. *Matrizentheorie*. Deutscher Verlag der Wissenschaften, Berlin, Germany. (A translation of the Russian original *Teoria Matriz*, Nauka Moscow, Russia, 1966.)
- GNEDENKO, B. V., AND KOLMOGOROV, A. N. 1968. *Limit Distributions for Sums of Independent Random Variables*. Addison-Wesley, Reading, MA.
- GUIBAS, L., AND ODLYZKO, A. M. 1981a. Periods in strings. *J. Combinat. Theory Ser. A*, 30, 19–43.
- GUIBAS, L., AND ODLYZKO, A. M. 1981b. String overlaps, pattern matching, and nontransitive games. *J. Combinat. Theory Ser. A*, 30, 183–208.
- GUIVARC'H, Y. 2000. Marches aléatoires sur les groupes. *Fascicule de probabilités*. Publ. Inst. Rech. Math. Rennes.
- GWADERA, R., ATALLAH, M., AND SZPANKOWSKI, W. 2004. Detection of significant sets of episodes in event sequences. In *Proceedings of the 4th IEEE International Conference on Data Mining (ICDM-04)* (Brighton, UK, Nov.) IEEE Computer Science Press, Los Alamitos, CA, 3–10.
- GWADERA, R., ATALLAH, M., AND SZPANKOWSKI, W. 2005. Markov models for identification of significant episodes. In *Proceedings of the 2005 SIAM International Conference on Data Mining (SDM-05)* (Newport Beach, Cal. Apr.). SIAM, Philadelphia, PA.
- HWANG, H.-K. 1996. Large deviations for combinatorial distributions: I. Central limit theorems, *Ann. Appl. Prob.* 6, 297–319.
- HWANG, H.-K. 1998. On convergence rates in the central limit theorems for combinatorial structures, *Europ. J. Combinat.* 19, 329–343.
- JANSON, S. 1997. *Gaussian Hilbert spaces*, Cambridge University Press, Cambridge, MA.
- JANSON, S. 2004. Large deviations for sums of partially dependent random variables. *Rand. Struct. Algorithms* 24, 234–248.
- KATO, T. 1980. *Perturbation theory for Linear Operators*. Springer-Verlag, New York.
- KNUTH, D. E. 1997. *The Art of Computer Programming, Fundamental Algorithms*, Vol. 1, 3rd Ed., Addison-Wesley, Reading, MA.
- KNUTH, D. E. 1998. *The Art of Computer Programming. Sorting and Searching*, Vol. 3, 2nd Ed., Addison-Wesley, Reading, MA.
- KUCHEROV, G., AND RUSINOWITCH, M. 1997. Matching a set of strings with variable length don't cares. *Theoret. Comput. Sci.* 178, 129–154.
- KUMAR, S., AND SPAFFORD, E. H. 1994. A pattern-matching model for intrusion detection. In *Proceedings of the National Computer Security Conference*. 11–21.
- LOTHAIRE, M. 2005. *Applied Combinatorics on Words*. In *Encyclopedia of Mathematics and Its Applications*, vol. 90. Cambridge University Press, Cambridge, MA.
- MCKAY, B. D., BAR-NATAN, D., BAR-HILLEL, M., AND KALAI, G. 1999. Solving the Bible code puzzle, *Statistical Science*, 14, 150–173.
- NICODÈME, P., SALVY, B., AND FLAJOLET, P. 1999. Motif statistics. In *Proceedings of the European Symposium on Algorithms*. IEEE Computer Society Press, Los Alamitos, CA, No. 1643. 194–211.
- PEVZNER, P. 2000. *Computational Molecular Biology: An Algorithmic Approach*, MIT Press, Cambridge, MA.
- RÉGNIER, M., AND SZPANKOWSKI, W. 1997. On the approximate pattern occurrences in a text. In *Proceedings of the Compression and Complexity of Sequences*, IEEE Computer Society Press, Los Alamitos, CA, 253–264.

- RÉGNIER, M., AND SZPANKOWSKI, W. 1998. On pattern frequency occurrences in a Markovian sequence. *Algorithmica* 22, 631–649.
- RIGOUTSOS, I., FLORATOS, A., PARIDA, L., GAO, Y., AND PLATT, D. 2000. The emergence of pattern discovery techniques in computational biology. *Metab. Eng.* 2, 159–177.
- SEDGEWICK, R., AND FLAJOLET, P. 1995. *An Introduction to the Analysis of Algorithms*, Addison-Wesley, Reading, MA.
- STEELE, J. M. 1997. *Probability Theory and Combinatorial Optimization*. SIAM, Philadelphia, PA.
- SZPANKOWSKI, W. 2001. *Average Case Analysis of Algorithms on Sequences*. Wiley, New York.
- VALLÉE, B. 2001. Dynamical sources in information theory: Fundamental intervals and word prefixes. *Algorithmica* 29, 262–306.
- VANET, A., MARSAN, L., AND SAGOT, M.-F. 1999. Promoter sequences and algorithmical methods for identifying them. *Res. Microbiol.* 150, 779–799.
- WATERMAN, M. 1995. *Introduction to Computational Biology*. Chapman and Hall, London, England.
- WESPI, A., DEBAR, H., DACIER, M., AND NASSEHI, M. 2000. Fixed vs. variable-length patterns for detecting suspicious process behavior. *J. Comput. Secur.* 8, 159–181.
- WITZTUM, D., RIPS, E., AND ROSENBERG, Y. 1994. Equidistant letter sequences in the book of genesis. *Stat. Sci.* 9, 429–438.
- WU, S., AND MANBER, U. 1995. Fast text searching allowing errors. *Commun. ACM* 35, 10 (Oct.), 83–91.

RECEIVED OCTOBER 2002; REVISED MAY 2005; ACCEPTED AUGUST 2005