# Classification of Markov Sources Through Joint String Complexity: Theory and Experiments

Philippe Jacquet and Dimitris Milioris
Bell Labs,
Alcatel-Lucent,
91620, France,
Email: {Philippe.Jacquet, dimitrios.milioris}@alcatel-lucent.com

Wojciech Szpankowski
Department of Computer Science,
Purdue University,
West Lafayette, IN 47907-2066 U.S.A.,
Email: spa@cs.purdue.edu

***Abstract*—We propose a classification test to discriminate Markov sources [19] based on the joint string complexity. String complexity is defined as the cardinality of a set of all distinct words (factors) of a given string. For two strings, we define *joint string complexity* as the set of words that are common to both strings. In this paper we analyze the average joint complexity when both strings are generated by a Markov source and provide fast converging asymptotic expansions. We also present some experimental results showing its usefulness to texts discrimination.**

## I. Introduction

In the last decades, several attempts have been made to capture mathematically the concept of "complexity" of a sequence, *i.e.* the number of different factors contained in a sequence. In other words, if $X$ is a sequence and $I(X)$ its set of factors (distinct subwords), then the cardinality $|I(X)|$ is the complexity of the sequence. For example, if $X = aabaa$ then $I(X) = \{\nu, a, b, aa, ab, aba, aab, abaa, aabaa\}$ and $|I(X)| = 9$ ($\nu$ denotes the empty string). Sometimes the sequence complexity is called the $I$-complexity of the sequence [5]. The notion is connected with quite deep mathematical properties, including the rather elusive concept of randomness in a string (see e.g., [3], [13], [14]).

In general, information contained in a string cannot be measured in absolute and a reference string is required. To this end we introduced in [4] the concept of *joint* complexity, or $J$-complexity, of two strings. The $J$-complexity is the number of different factors common to two sequences. In other words the $J$ complexity of sequence $X$ and $Y$ is equal to $J(X, Y) = |I(X) \cap I(Y)|$. We denote $J_{n,m}$ the average value of $J(X, Y)$ when length of $X$ is $n$ and length of $Y$ is $m$. In this paper we study in this paper its growth when $n = m$.

The $J$-complexity is an efficient way of evaluating similarity degree of two sequences. For example, genome sequences of two dogs will contain more common words than genome sequences of a dog and a cat. Similarly, two texts written in the same language have more common words than texts written in very different languages. Also the $J$-complexity is larger when languages are close (*e.g.* French and Italian), than when languages are very different (*e.g.* French and English). Furthermore, texts in the same language but on different topics (*e.g.* Law and cooking) have smaller $J$ complexity than texts on the same topic (*e.g.* medicine). Therefore the $J$-complexity is a pertinent tool for automated monitoring of social networks. But for this purpose the $J$-complexity should discriminate well short texts. This requires a precise analysis of the joint complexity, which we offer in this paper (see also [8]) together with some experimental results (cf. Figures 1 and 2) confirming usefulness of the joint string complexity to texts discrimination.

In [4] it is proved that the $J$-complexity of two texts built from two *different* binary memoryless sources grows like $\gamma \frac{n^\kappa}{\sqrt{\alpha \log n}}$, for some $\kappa < 1$ and $\gamma, \alpha > 0$ which depend on the parameters of two sources. When the sources are identical, then the $J$-complexity growth is $O(n)$, hence $\kappa = 1$. When the texts are identical (i.e, $X = Y$), then the $J$-complexity is identical to the $I$-complexity and it grows as $\frac{n^2}{2}$ [11]. Therefore the $J$-complexity can already be used to detect "copy-paste" between texts; indeed the presence of a common factor of length $O(n)$ would inflate the $J$ complexity by a term $O(n^2)$.

We should point out that experiments show that the complexity estimate as above for memoryless sources converges very slowly. Therefore, joint complexity is not really meaningful even when $n \approx 10^9$. Furthermore, memoryless sources are not appropriate for modeling text generation. In this paper we extend the $J$-complexity estimate to Markov sources of any order on a finite alphabet. Although Markov models are no more realistic, say for DNA sequence, than memoryless sources, but for text generation it seems to be fairly realistic.

In this paper we derive a second order asymptotics for $J$-complexity of the following form $\gamma \frac{n^\kappa}{\sqrt{\alpha \log n + \beta}}$ . for some $\beta > 0$. This new estimate converges more quickly, and usually works for texts of order $n \approx 10^2$; thus it can be used for short text such as tweets. In fact, for some Markov sources our analysis indicate that $J$-complexity oscillates with $n$. This manifestates in the factor $P\left(\frac{1}{\alpha \log n + \beta}\right)$ appearing in the $J$ complexity, where $P$ is a specific polynomial determined via saddle point expansion. This additional term even further improves the convergence for small values of $n$.

In view of these facts, we can use the $J$ complexity to discriminate between two identical/non-identical Markov sources

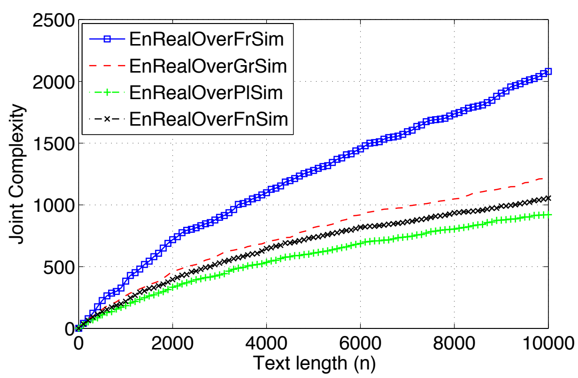


Fig. 1. Joint complexity of an English text vs French, Greek, Polish, and Finnish texts.

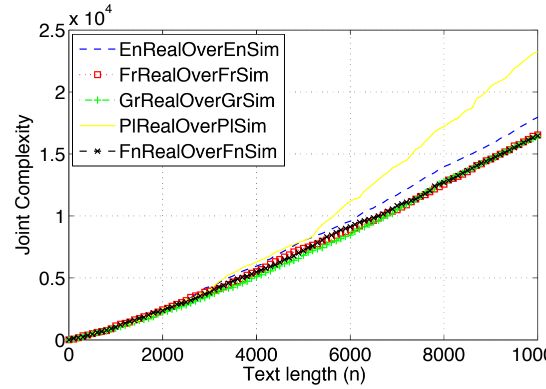


Fig. 2. Joint complexity of real and simulated texts (3rd Markov order) of English, French, Greek, Polish and Finnish language.

[19]. We introduce the discriminant function as follows

$$d(X,Y) = 1 - \frac{1}{\log n} \log J(X,Y)$$

for two sequences $X$ and $Y$ of length $n$. This discriminant allows us to determine whether $X$ and $Y$ are generated by the same Markov source or not by verifying whether $d(X,Y) = O(1/\log n) \to 0$ or $d(X,Y) = 1 - \kappa + O(\log\log n/\log n) > 0$, respectively when the length of $X$ and $Y$ are both equal to $n$. In this conference paper we mainly concentrate on the analysis of $J$-complexity leaving further analysis of the discriminant $d(X,Y)$ to a forthcoming full paper. However, we present below some experimental evidence of how useful our discriminant is for real texts.

In Figure 1 we compared the joint complexity of an English text to the same length texts written in French, Greek, Polish and Finnish. It is easy to see that even for texts of lengths smaller than a thousand one can discriminate between these languages. On the other hand, in Figure 2 we plot the joint complexity between real and simulated texts in French, Greek, Polish, English and Finnish. Clearly, the joint complexity of such texts grows like $O(n)$ as predicted by theory. In fact, computations show that with Markov models of order 3 for English versus French we have $\kappa = 0.44$; versus Greek: $\kappa = 0.26$; versus Finnish $\kappa = 0.04$; and versus Polish: $\kappa = 0.01$, which is consistent with the results on Figure 1. (In fact, they agree with theoretical results of Theorem 3 discussed below.)

Single string complexity was studied extensively in the past. The literature is reviewed in [11] where precise analysis of string complexity is discussed for strings generated by unbiased memoryless sources. Another analysis of the same situation was also proposed in [4] where for the first time the joint string complexity for memoryless sources is presented. It was evident from [4] that precise analysis of the joint complexity is quite challenging due to intricate singularity analysis and infinite number of saddle points. In this paper we deal with the joint string complexity for Markov sources. To the best of our knowledge this problem was never tackled before. As expected, its analysis is very sophisticated but at the same time quite rewarding. It requires generalized (two-dimensional) depoissonization and generalized (two-dimensional) Mellin transforms. In fact, we discovered new oscillation phenomena.

## II. Main Results

### A. Models and notations

We begin by introducing some general notation. Let $\omega$ and $\sigma$ be two strings over alphabet $\mathcal{A}$. We denote by $|\omega|_\sigma$ the number of times $\sigma$ occurs in $\omega$ (e.g., $|abbba|_{bb} = 2$). By convention $|\omega|_\nu = |\omega| + 1$, where $\nu$ is the empty string.

Throughout we denote by $X$ a string (text) whose complexity we plan to study. We also assume that its length $|X|$ is equal to $n$. Then the string complexity is $I(X) = \{\omega : \ |X|_\omega \geq 1\}$. Observe that

$$|I(X)| = \sum_{\sigma \in \mathcal{A}^*} 1_{|X|_\sigma \geq 1},$$

where $1_A$ is the indicator function of a boolean $A$. Notice that $|I(X)|$ is equal to the number of nodes in the associated suffix tree of $X$ [18] (see also [6]).

Now, let $X$ and $Y$ be two sequences (not necessarily of the same length). We define the *joint complexity* as the cardinality of the set $J(X,Y) = I(X) \cap I(Y)$. The joint semi-complexity is the cardinality of the set $S(X,Y) = I_2(X) \cap I_2(Y)$ In fact, the joint semi-complexity corresponds to the number of common nodes in two suffix trees built from $X$ and $Y$. We have

$$|J(X,Y)| = \sum_{\sigma \in \mathcal{A}^*} 1_{|X|_\sigma \geq 1} \times 1_{|Y|_\sigma \geq 1} \ .$$

We now assume that both strings $X$ and $Y$ are generated by two *independent Markov sources* of order $r$ (we will only detail the analysis for order 1, but extension to arbirary order is straightforward). We assume that source $i$, for $i \in \{1,2\}$ has the transition probabilities $P_i(a|b)$ from state $b$ to state $a$, where $(a,b) \in \mathcal{A}^r$. We denote by $\mathbf{P}_1$ (resp. $\mathbf{P}_2$) the transition matrix of Markov source 1 (resp. source 2). The stationary distributions are respectively denoted by $\pi_1(a)$ and $\pi_2(a)$ for $a \in \mathcal{A}^r$.

Let $X_n$ and $Y_n$ be two strings of respective length $n$ and $m$, $X_n$ generated on Markov source 1, and $Y_m$ generated on Markov source 2. We write $J_{n,m} = \mathbf{E}(|J(X_n,Y_n)|) - 1$ for the joint complexity, *i.e.* omitting the empty string.

### B. Summary of Main Results

We say that a matrix $\mathbf{M} = [m_{ab}]_{(a,b)\in\mathcal{A}^2}$ is *rationally balanced* if $\forall(a,b,c) \in \mathcal{A}^2$: $m_{ab} + m_{ca} - m_{cb} \in \mathbb{Z}$. We say that a positive matrix $\mathbf{M} = [m_{ab}]$ is *logarithmically rationally balanced* when the matrix $\log^*(\mathbf{M}) = [l_{ab}]$ where $\ell_{ab} = \log(m_{ab})$ when $m_{ab} > 0$ and $l_{ab} = 0$ otherwise. Furthermore, we say that two matrices $\mathbf{M} = [m_{ab}]_{(a,b)\in\mathcal{A}^2}$ and $\mathbf{M}' = [m'_{ab}]$ are *logarithmically commensurable* when matrices $\log^*(\mathbf{M})$ and $\log^*(\mathbf{M}')$ are commensurable. That is, there exist a nonzero pair of reals $(x, y)$ such that $x\log^*(\mathbf{M}) + y\log^*(\mathbf{M}')$ is logarithmically rationally balanced.

We now present our main theoretical results in a series of theorems each treating different cases of Markov sources.

*Theorem 1:* Consider the average joint complexity of two texts of length $n$ generated by the same general stationary Markov source.
(i) [*Noncommensurable Case.*] Assume that $\mathbf{P}_1 = \mathbf{P}_2$ are not logarithmically rationally balanced. Then

$$J_{n,n} = \frac{2\log 2}{h} n + o(1) \tag{1}$$

where $h$ is the entropy rate of the source.
(ii) [*Commensurable Case.*] Assume that $\mathbf{P}_1 = \mathbf{P}_2$ are logarithmically rationally balanced. Then there are periodic terms and $\epsilon > 0$ such that: $J_{n,n} = \frac{2\log 2}{h}(1 + Q_0(\log n)) + O(n^{-\epsilon})$ where $Q_0(.)$ is a periodic function of small amplitude.

Now we consider sources that are not the same and have respective transition matrices $\mathbf{P}_1$ and $\mathbf{P}_2$. The transition matrices are on $\mathcal{A}^r \times \mathcal{A}^r$ where $r$ denotes the order of the underlying Markov sources. If $(a,b) \in \mathcal{A}^r \times \mathcal{A}^r$, we denote by $\mathbf{P}_i(a|b)$ the $(a,b)$ coefficient of matrix $\mathbf{P}_i$. For a tuple of complex numbers $(s_1, s_2)$ we define $\mathbf{P}(s_1, s_2)$ as the matrix whose $(a,b)$ coefficient is $(\mathbf{P}_1(a|b))^{-s_1}(\mathbf{P}_2(a|b))^{-s_2}$.

We first consider the case when matrix $\mathbf{P}(s_1, s_2)$ is nilpotent [12].

*Theorem 2:* If $\mathbf{P}(s_1, s_2)$ is nilpotent, then there exists $\gamma_0$ and $\epsilon > 0$ such that $\lim_{n\to\infty} J_{n,n} = \gamma_0$.

This result is not surprising and rather trivial since the common factors can only occur in a finite window at the beginning of the strings.

Throughout, now we assume that $\mathbf{P}(s_1, s_2)$ is not nilpotent. We denote by $\mathcal{K}$ the set of real tuple $(s_1, s_2)$ such that $\mathbf{P}(s_1, s_2)$ has the main eigenvalue equal to 1. Let

$$\begin{aligned}
\kappa &= \min_{(s_1,s_2)\in\mathcal{K}} \{-s_1 - s_2\} \\
(c_1, c_2) &= \arg\min_{(s_1,s_2)\in\mathcal{K}} \{-s_1 - s_2\}.
\end{aligned}$$

Easy algebra proves that $\kappa < 1$.

*Theorem 3:* Assume $\mathbf{P}(s_1, s_2)$ is not nilpotent and either $c_1 > 0$ or $c_2 > 0$.
(i) [*Noncommensurable Case.*] We assume that $\mathbf{P}_2$ is noncommensurable. Let $c_0 < 0$ such that $(c_0, 0) \in \mathcal{K}$. There exists $\gamma_1$ and $\epsilon > 0$:

$$J_{n,n} = \gamma_1 n^{-c_0}(1 + O(n^{-\epsilon})) \tag{2}$$

(ii) [*Commensurable Case.*] Let now $\mathbf{P}_2$ be logarithmically rationally balanced. There exists a periodic function $Q_1(.)$ of small amplitude such that $J_{n,n} = \gamma_1 n^{-c_0}(1 + Q_1(\log n) + O(n^{-\epsilon})$.

The case where both $c_1$ and $c_2$ are between $-1$ and 0 is the most intricate to handle. We summarize our results next.

*Theorem 4:* Assume that $c_1$ and $c_2$ are between $-1$ and 0.
(i) [*Noncommensurable Case.*] When $\mathbf{P}_1$ and $\mathbf{P}_2$ are noncommensurable, then there exists $\alpha_2$, $\beta_2$ and $\gamma_2$ such that

$$J_{n,n} = \frac{\gamma_2 n^\kappa}{\sqrt{\alpha_2 \log n + \beta_2}}(1 + o(1)) \ . \tag{3}$$

(ii) [*Commensurable Case.*] Let $\mathbf{P}_1$ and $\mathbf{P}_2$ be logarithmically commensurable matrices. Then there exist a double periodic function $Q_2(.)$ of small amplitude such that:

$$J_{n,n} = \frac{\gamma_2 n^\kappa}{\sqrt{\alpha_2 \log n + \beta_2}}(1 + Q_2(\log n) + o(1)).$$

## III. Theoretical Analysis

In this section we present a sketch of ideas that proves our main results. The technique we use in fact allows us to prove much stronger (refined) results presented at the end of this section.

### A. Equivalence Suffixes and independent strings

We have the identity:

$$J_{n,m} = \sum_{w\in\mathcal{A}^*-\{\nu\}} P(w \in I(X_n)) \times P(w \in I(X_n) \geq 1) \ . \tag{4}$$

We know that from [6], [15] that there is a close formula for $\sum_n P(|X_n|_w \geq 1)z^n = \frac{P_1(w)z}{(1-z)D_w(z)}$ which is, in the memoryless case.

$$D_w(z) = (1-z)(1 + A_w(z)) + P_1(w)z^{|w|} \ , \tag{5}$$

where $P(w)$ is the probability that $w$ is prefix of $X_n$, and $A_w(z)$ is the *autocorrelation* polynomial of word $w$. For the Markov source, we omit the expression which carries extra indices to keep track with the Markov correlations with the starting symbols of the words (for a complete description of the parameters see [6], [15]).

Although being a closed formula, this expression is not easy to manipulate. To make the analysis tractable we notice that $w \in I(X_n)$ is equivalent to the fact that $w$ is at least prefix of one of the $n$ suffices of $X_n$. If the suffices would have been $n$ independent infinite strings then $P(w \in I(X_n))$ would be equal to $1 - (1 - P_1(w))^n$ whose generating function is $\frac{P_1(z)z}{(1-z)(1-z+P_1(w)z)}$, which rather the same as $\frac{P_1(w)z}{(1-z)D_w(z)}$ but with $A_w(z) = 0$ and $z^{|w|} = z$.

We define $I_1(n)$ (resp. $I_2(n)$) as the set of prefixes of $n$ independent strings built on source 1 (resp. 2). Let

$$C_{n,m} = \sum_{w\in\mathcal{A}^*-\{\nu\}} P(w \in I_1(n))P(w \in I_2(n)) \ .$$

Using the techniques of [6], [15]) we prove that

*Lemma 1:* For some $\epsilon > 0$

$$J_{n,n} = C_{n,n}\left(1 + O(n^{-\epsilon})\right) + O(1). \tag{6}$$

A proof of this lemma follows from [6], [15], and will be given in the journal version of this paper.

### B. Functional equations

Let $a \in \mathcal{A}$. We denote $C_{a,m,m}$ the quantity $\sum_{w \in a\mathcal{A}^*} P(w \in I_1(n))P(w \in I_2(n))$. Notice that $C_{a,m,n} = 0$ when $n = 0$ or $m = 0$. Using the Markov nature of the string generation, the quantity $C_{a,n,m}$ for $n, m \geq 1$ satisfies the following recurrence for all $b \in \mathcal{A}$

$$\begin{aligned} C_{b,n,m} &= 1 + \sum_{a \in \mathcal{A}} \sum_{n_a, m_a} \binom{n}{n_a}\binom{m}{m_a} \\ &\times (P_1(a|b))^{n_a}(1 - P_1(a|b))^{n-n_a} \\ &\times (P_2(a|b))^{m_a}(1 - P_2(a|b))^{m-m_a} C_{a,n_a,m_a} \ , \end{aligned}$$

where $n_a$ denotes the number of strings among the $n$ independent strings on source 1 which have symbol $a$ that follows symbol $b$ as second character. Quantity $m_a$ is the counterpart on source 2. The *unconditional* average $C_{n,m}$ satisfies for $n, m \geq 2$

$$\begin{aligned} C_{n,m} &= 1 + \sum_{a \in \mathcal{A}} \sum_{n_a, m_a} \binom{n}{n_a}\binom{m}{m_a} \pi_1^{n_a}(a)(1 - \pi_1(a))^{n-n_a} \\ &\times \pi_2^{m_a}(a)(1 - \pi_2(a))^{m-m_a} C_{a,n,m}. \end{aligned}$$

We introduce the double Poisson transform of $C_{a,n,m}$

$$C_a(z_1, z_2) = \sum_{n,m \geq 0} C_{a,n,m} \frac{z_1^n z_2^m}{n!m!} e^{-z_1 - z_2} \tag{7}$$

translates the above recurrence into the following functional equation:

$$\begin{aligned} C_b(z_1, z_2) &= (1 - e^{-z_1})(1 - e^{-z_2}) \\ &+ \sum_{a \in \mathcal{A}} C_a\left(P_1(a|b)z_1, P_2(a|b)z_2\right). \end{aligned} \tag{8}$$

Furthermore, the cumulative double Poisson transform

$$C(z_1, z_2) = \sum_{n,m \geq 0} T_{n,m} \frac{z_1^n z_2^m}{n!m!} e^{-z_1 - z_2} \tag{9}$$

which satisfies

$$\begin{aligned} C(z_1, z_2) &= (1 - e^{-z_1})(1 - e^{-z_2}) \\ &+ \sum_{a \in \mathcal{A}} C_a(\pi_1(a)z_1, \pi_2(a)z_2) \ . \end{aligned} \tag{10}$$

### C. DePoissonization

Using [7], [8], [18] we prove

*Lemma 2 (DePoissonization):* When $n$ and $m$ tend to infinity:

$$C_{n,m} = C(n, m)(1 + O(\frac{1}{n}) + O(\frac{1}{m})) \ .$$

This equivalence is obtained by proving some growth properties of $C(z_1, z_2)$ when $(z_1, z_2)$ are complex numbers.

### D. Same Markov sources

We first give a general result when the Markov sources are identical: $\mathbf{P}_1 = \mathbf{P}_2 = \mathbf{P}$. In this case equation 8 can be rewritten with $c_a(z) = C_a(z, z)$:

$$c_b(z) = (1 - e^{-z})^2 + \sum_{a \in \mathcal{A}} c_a\left(P(a|b)z\right). \tag{11}$$

This equation is directly solvable by introducing the Mellin transform $c_a^*(s) = \int_0^\infty c_a(x)x^{s-1}dx$ defined for $-2 < \Re(s) < -1$ and which satisfies the equation for all $b \in \mathcal{A}$.

$$c_b^*(s) = (2^{-s} - 2)\Gamma(s) + \sum_{a \in \mathcal{A}} (P(a|b))^{-s} c_a^*(s) \ . \tag{12}$$

Introducing $c^*(s)$ the Mellin transform of $C(z, z)$ we get the identity:

$$c^*(s) = (2^{-s} - 2)\Gamma(s) + \sum_{a \in \mathcal{A}} (\pi(a))^{-s} c_a^*(s) \ .$$

Thus

$$c^*(s) = (2^{-s} - 2)\Gamma(s)\left(1 + \langle \mathbf{1}(\mathbf{I} - \mathbf{P}(s))^{-1} | \boldsymbol{\pi}(s)\rangle\right) \tag{13}$$

where $\mathbf{1}$ is the vector of dimension $|\mathcal{A}|$ made of 1's, $\mathbf{I}$ is the identity matrix, and $\mathbf{P}(s) = \mathbf{P}(s, 0) = \mathbf{P}(0, s)$, $\boldsymbol{\pi}(s)$ is the the vector made of coefficients $\pi(a)^{-s}$ and $\langle .|. \rangle$ denotes the inner product.

By applying the methodology of Flajolet [2], [18], the asymptotics of $c(z)$ for $|\arg(z)| < \theta$ is given by the residues of the function $c^*(s)z^{-s}$ which occurs on $s = -1$ and $s = 0$ and which are respectively $\frac{2\log 2}{h}z$ and $-1 - \langle \mathbf{1}(\mathbf{I} - \mathbf{P}(0, 0))^{-1}\boldsymbol{\pi}(0)\rangle$. The first residues comes from the singularity of $(\mathbf{I} - \mathbf{P}(s))^{-1}$ on $s = -1$. This lead to the formula of Theorem 4. When $\mathbf{P}$ is logarithmically rationally balanced then there are additional poles on a countable set of complex numbers $s_k$ regularly spaced on the same imaginary axes containing $-1$ and such that $\mathbf{P}(s_k)$ has eigenvalue 1. These poles contributes to the periodic terms of Theorem 4.

Computations show that a Markov model of order 3 for English text has entropy: 0.944221; French entropy s 0.934681; Greek: 1.013384, Polish: 0.665113; and Finnish entropy is 0.955442. This is consistent with Figure 2

### E. Different Markov sources

In this section we identify the constants in Theorems 3 and 4.

Since $\mathbf{P}_1 \neq \mathbf{P}_2$ we cannot get a functional equation for the $C_a(z, z)$'s, we thus have to deal with the two variables $z_1$ and $z_2$. We define the double Mellin transform $C_a^*(s_1, s_2) = \int_0^\infty \int_0^\infty C_a(z_1, z_2) z_1^{s_1-1} z_2^{s_2-1} dz_1 dz_2$ and simiarly $C^*(s_1, s_2)$ the double Mellin transform of $C(z_1, z_2)$. And thus we have the identity

$$\begin{aligned} C_b^*(s_1, s_2) &= \Gamma(s_1)\Gamma(s_2) \\ &+ \sum_{a \in \mathcal{A}} (P_1(a|b))^{-s_1}(P_2(a|b))^{-s_2} C_a^*(s_1, s_2) \end{aligned} \tag{14}$$

and

$$C^*(s_1,s_2)=\Gamma(s_1)\Gamma(s_2)\left(1+\langle \mathbf{1}(\mathbf{I}-\mathbf{P}(s_1,s_2))^{-1}|\boldsymbol{\pi}(s_1,s_2)\rangle\right) \quad (15)$$

where $\boldsymbol{\pi}(s_1,s_2)$ denotes the vector made of coefficients $\pi_1(a)^{-s_1}\pi_2(a)^{-s_2}$. In fact to be defined the Mellin transform we need to apply it on $C(z_1,z_2)-\frac{\partial}{\partial z_1}C(0,z_2)z_1e^{-z_1}-\frac{\partial}{\partial z_2}C(z_1,0)z_2e^{-z_2}$ but we omit this technical detail.

The inverse Mellin transform is

$$C(z,z)=\frac{1}{(2i\pi)^2}\int_{\Re(s_1)=\rho_1}\int_{\Re(s_2)=\rho_2}C^*(s_1,s_2)z^{-s_1-s_2}ds_1ds_2 \quad (16)$$

where $(\rho_1,\rho_2)$ belongs to the definition domain of $C^*(s_1,s_2)$.

We denote $L(s)$ the function of complex $s$ such that $\mathbf{P}(s,L(s))$ has eigenvalue 1. The function is meromorphic and has several branches; one branches describes the set $\mathcal{K}$ when $s$ is real. Via the formula of residues we can get rid of variable $s_2$ by letting $\rho_2$ moving to some non negative value $M$:

$$\begin{aligned}C(z,z) &= \frac{1}{2i\pi}\int_{\Re(s_1)=\rho_1}\mu(s_1)\Gamma(s_1)\Gamma(L(s_1))z^{-s_1-L(s_1)}ds_1\\ &\quad +O(z^{\rho_1-M})\end{aligned}$$

where $\mu(s)$ is the residue of function $\langle\mathbf{1}(\mathbf{I}-\mathbf{P}(s,s_2))^{-1}\boldsymbol{\pi}(s_1,s_2)\rangle$ at point $(s,L(s))$, actually it is equal to $\frac{1}{\frac{\partial}{\partial s_2}\lambda(s_1,s_2)}\langle\mathbf{1}|\boldsymbol{\zeta}(s_1,s_2)\rangle\langle\mathbf{u}(s_1,s_2)|\boldsymbol{\pi}(s_1,s_2)\rangle$, where $\lambda(s_1,s_2)$ is the eigenvalue which has value 1 at $(s,L(s))$ and $\mathbf{u}(s_1,s_2)$ and $\boldsymbol{\zeta}(s_1,s_2)$ are the respective left and right eigenvector with the convention that $\langle\boldsymbol{\zeta}(s_1,s_2)|\mathbf{u}(s_1,s_2)\rangle=1$.

The expression is implicitly a sum since the function $L(s)$ is meromorphic, but we retain only the branch where $\lambda(s_1,s_2)$ is the main eigenvalue of $\mathbf{P}(s_1,s_2)$ for the leading term in the expansion of $C(z,z)$. For more details see [8] where the analysis is detailled in the case where $\mathbf{P}_2$ corresponds to the uniform memoryless case, *i.e.* $\mathbf{P}_2=\frac{1}{|\mathcal{A}|}\mathbf{1}\otimes\mathbf{1}$.

The point here is to move the integration line for $s_1$ from $\rho_1$ to $c_1$ which corresponds to the position the minimum of function $-s_1-L(s_1)$ (actually $\kappa$). We only consider the case where $L(c_1)=c_2<0$ (the other case is obtained by symmetry). The poles are due to the function $\Gamma(.)$. The first encountered pole is $s_1=-1$ but this pole cancels with the technical arrangement discussed earlier.

Let assume $c_1>0$, therefore we meet a second pole on $s=0$ and the residue, equal to $\mu(0)\Gamma(c_0)z^{-c_0}$ since $L(0)=c_0$. This quantity turns out to be the leading term of $C(z,z)$ since the integration on $\Re(s_1)=c_1$ is in $O(z^\kappa)$. This prove theorem 3. In the case $\mathbf{P}_2$ is commensurable, there exists $\nu$ such that $\lambda(s,L(s)+ik\nu)=1$ and therefore terms in $z^{c_0+ik\nu}$ give a periodic contribution.

The most tricky part is when $-1<c_1<0$. In this case we get an estimate in $O(z^\kappa)$ but to get precise estimate one must use of the saddle point methods on $s=c_1$ since the integration is of the form $\int_{\Re(s)=c_1}\mu(s)\exp(-(s+L(s))A)ds$ and $A=\log z\to\infty$. We naturally get an expansion

$$C(z,z)=\frac{e^{\kappa A}\mu(c_1)}{\sqrt{(\alpha_2A+\beta_2}}\left(1+O(\frac{1}{\sqrt{A}}\right)$$

with $\alpha_2=L''(c_1)$ and $\beta_2=\frac{\mu'(c_1)}{\mu(c_1)}$. In fact the saddle point expansion is extendable to any oder of $\frac{1}{\sqrt{A}}$. This proves the theorem 4 in the general case. For the case where $\mathbf{P}_1$ and $\mathbf{P}_2$ are logarithmically commensurable, the line $\Re(s_1)=c_1$ contains an infinite number of saddle points that contribute in a double periodic additional term.

## ACKNOWLEDGMENT

W. Szpankowski work was partially supported by the NSF Science and Technology Center for Science of Information Grant CCF-0939370, NSF Grants DMS-0800568 and CCF-0830140, and, NSA Grant H98230-11-1-0141. D. Milioris work is also supported by INRIA and Ecole Polytechnique Doctorale school.

## REFERENCES

[1] P. Flajolet, X. Gourdon, and P. Dumas, Mellin Transforms and Asymptotics: Harmonic sums, *Theoretical Computer Science*, 144, 3–58, 1995.
[2] P. Flajolet and R. Sedgewick, *Analytic Combinatorics*, Cambridge University Press, Cambridge, 2008.
[3] Ilie, L., Yu, S., and Zhang, K. Repetition Complexity of Words In *Proc. COCOON* 320–329, 2002.
[4] P. Jacquet, Common words between two random strings, *IEEE Intl. Symposium on Information Theory*, 1495-1499, 2007.
[5] V. Becher and P. A. Heiber, A better complexity of finite sequences, Abstracts of the 8th *Int. Conf. on Computability and Complexity in Analysis* and 6th *Int. Conf. on Computability, Complexity, and Randomness* , Cape Town, South Africa, January 31, February 4, 2011, p. 7.
[6] P. Jacquet, and W. Szpankowski, Autocorrelation on Words and Its Applications. Analysis of Suffix Trees by String-Ruler Approach, *J. Combinatorial Theory Ser. A*, 66, 237–269, 1994.
[7] P. Jacquet, and W. Szpankowski, Analytical Depoissonization and Its Applications, *Theoretical Computer Science*, 201, 1–62, 1998.
[8] P. Jacquet and W. Szpankowski, Joint String Complexity for Markov Sources, *23rd International Meeting on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms*, AofA'12, *DMTCS Proc.*, 303-322, Montreal, 2012.
[9] P. Jacquet, and W. Szpankowski, Analytic Approach to Pattern Matching, Chap. 7 in *Applied Combinatorics on Words* (eds. Lothaire), Cambridge University Press (Encycl. of Mathematics and Its Applications), Cambridge, 2005.
[10] P. Jacquet, W. Szpankowski, and J. Tang, Average Profile of the Lempel-Ziv Parsing Scheme for a Markovian Source, *Algorithmica*, 31, 318-360, 2001.
[11] S. Janson, S. Lonardi and W. Szpankowski, On Average Sequence Complexity, *Theoretical Computer Science*, 326, 213-227, 2004.
[12] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, 1985.
[13] Li, M., and Vitanyi, P. *Introduction to Kolmogorov Complexity and its Applications.* Springer-Verlag, Berlin, Aug. 1993.
[14] Niederreiter, H., Some computable complexity measures for binary sequences, In *Sequences and Their Applications*, Eds. C. Ding, T. Hellseth and H. Niederreiter Springer Verlag, 67-78, 1999.
[15] J. FAYOLLE, M. WARD Analysis of the average depth in a suffix tree under a Markov model DMTCS Proceedings of AofA 2005.
[16] J. FAYOLLE Compression de données sans perte et combinatoire analytique Thèse de l'Université de Paris 6, 2006.
[17] G. Park, H.K. Hwang, P. Nicodeme, and W. Szpankowski, Profile of Tries, *SIAM J. Computing*, 8, 1821-1880, 2009.
[18] W. Szpankowski, *Analysis of Algorithms on Sequences*, John Wiley, New York, 2001.
[19] J. Ziv, On classification with empirically observed statistics and universal data compression, *IEEE Trans. Information Theory*, 34, 278-286, 1988.