AofA 2014, Paris, France

Information in Strings: Enumeration of Eulerian Paths and Eulerian Components in Markov Sequences

Philippe Jacquet[†] and Dimitrios Milioris[‡]

Bell Labs, Alcatel-Lucent
Nozay, France

Abstract. In this paper, we evaluate the number of Eulerian circuits that can be obtained by an arbitrary rotation in a Markovian string, *i.e.*, corresponding to a given Markovian type. Since all rotations do not result in an Eulerian circuit, but several of them, called Eulerian components; we also investigate the number of Eulerian components that result from a random rotation in a Markovian string. We consider the asymptotic behaviour of those quantities when the size of the string n tends to infinity. In particular we show that the average number of components tends to be in $\log V$, where V is the size of a large alphabet, in the uniform case.

Keywords: Asymptotic analysis, Eulerian paths and circuits, Analytic combinatorics

1 Introduction

In this paper we consider X_n, a string of length n generated by a Markov process over an alphabet $\mathcal{A}$ of size V. We denote $K(\mathcal{A})$ the set of integer $V \times V$ matrices defined on $\mathcal{A} \times \mathcal{A}$. Let $\mathbf{k} \in K(\mathcal{A})$ when $\forall (a,b) \in \mathcal{A}^2$ the coefficient k_{ab} of $\mathbf{k}$ is equal to the number of time symbol b follows symbol a in X_n, we say that $\mathbf{k}$ is the *type* of string X_n as defined in [5]. The matrix $\mathbf{k}$ can be seen as the adjacency matrix of the multigraph $G(X_n)$ whose vertex set is $\mathcal{A}$ and edge set collects all the transitions between consecutive symbols in X_n.

The string X_n is an Eulerian path in $G(X_n)$, *i.e.*, a path that takes every edge of $G(X_n)$ once and only once. Let $a \in \mathcal{A}$ and let $k_a = \sum_{b \in \mathcal{A}} k_{ab}$. An Eulerian path on X_n will visit symbol a, k_a times, and every time takes one outgoing edge among the non already visited edge, until all edges have been visited. The order at which the edge are visited is a permutation σ_a of the outgoing edges of symbol a. There are $\binom{k_a}{(k_{ab})_{b \in \mathcal{A}}} = \frac{k_a!}{\prod_{b \in \mathcal{A}} k_{ab}!}$, since we don't distinguish over the edges that connect to the same next symbol. Therefore an Eulerian path is defined by an initial symbol and a V tuple $(\sigma_a)_{a \in \mathcal{A}}$ of such permutations

[†]Partially supported by LINCS
[‡]Partially supported by INRIA and École Polytechnique

that we call *rotation*. There are $B_{\mathbf{k}}$ rotations with

$$B_{\mathbf{k}} = \prod_{a \in \mathcal{A}} \binom{k_a}{(k_{ab})_{b \in \mathcal{A}}}. \tag{1}$$

Of course not every rotation in $G(X_n)$ will generate an Eulerian path, since the path may end on a symbol which has exhausted all its outgoing edges before some other symbols exhaust their. We denote $N_{\mathbf{k}}^a$ the number of Eulerian path that shares the same type $\mathbf{k}$ starting from symbol a, and we naturally have $N_{\mathbf{k}}^a \leq B_{\mathbf{k}}$. Notice that all these Eulerian paths corresponds to strings that have the same probability in the Markov model.

Let R_n be the average number of rotation for a string a length n generated by the Markov process. We show that $R_n \sim A\lambda^{2n}$ for some $A > 0$ and $\lambda > 1$ that depends on the transition matrix. This quantity is almost impossible to simulate when n is large since the main contributions come from types whose probability is exponentially small.

Let L_n denote the average of the logarithm of the number of rotations, we show that in general $L_n \neq \log R_n$ and that in fact we have $L_n \sim H(X_n)$ where $H(X_n)$ is the entropy of the Markov string X_n.

When closing the string X_n we make it cyclical. In this case it is equivalent to add an edge in $G(X_n)$ between the last and the first symbol of X_n, and we get the new graph $\tilde{G}(X_n)$. We show such a graph in Figure 1 for the sentence "to be or not to be that is the question". We call *anchored* Eulerian circuit when we specify the initial symbol. Several anchored Eulerian circuits may correspond to the same Eulerian circuit, just translated by circular permutations. An anchored Eulerian path corresponds to a rotation in $\tilde{G}(X_n)$ but contrary to Eulerian path it may not have its symbol before last equal to the last symbol of X_n, even if the first symbols are the same, thus an anchored Eulerian circuit in $\tilde{G}(X_n)$ is not equivalent to an Eulerian path in $G(X_n)$. To make it equivalent one may consider an anchored Eulerian circuit starting on the last symbol of X_n, say symbol b, and force the first edge permutation to start with the edge ba; a being the first symbol of X_n.

When a rotation in $\tilde{G}(X_n)$ is not an Eulerian circuit, it can be decomposed into a sequence of paths in $\tilde{G}(X_n)$ that are edge-disjoint, that we call *Eulerian components*. We will provide an asymptotic estimate of the average number of such components. In particular when the Markov process is memoryless and uniform the probability that a rotation is an Eulerian circuit is asymptotically equal to $\frac{1}{V}$ and the average number of component is asymptotically equal to $H(V)$, the harmonic function of V.

The paper is organized as follows: Section 2 investigates the enumeration of Eulerian paths and other parameters via generating functions, while Section 3 applies these results on a Markov generating process and gives an analysis of rotations. Finally, Section 4 presents some experimental results and gives conclusions.

2 Balanced types, generating functions and Eulerian path

Pioneering work on Eulerian paths and circuits enumeration could be found in [1, 2] but we rely and extend the work and methodology of [5]. Notice the related works in [3] about Eulerian circuits in complete graphs and [4] about Eulerian path in bipartite graphs.

2.1 Balanced types

Let $c \in \mathcal{A}$ and a type $\mathbf{k}$ we denote $k_c = \sum_{d \in \mathcal{A}} k_{cd}$ and $k^c = \sum_{d \in \mathcal{A}} k_{dc}$ respectively the out-degree and in-degree of symbol c in the syntax graph. A type is balanced if $\forall a \in \mathcal{A}$: $k_a = k^a$. Let $\mathcal{F}_n$ be the set

balanced types $\mathbf{k}$ such that

- $k_\alpha > 0$;
- $\sum_{a\in\mathcal{A}} k_a = n$.

We also denote $\mathcal{F}_* = \bigcup_n \mathcal{F}_n$ as in [5]. If $\mathcal{C} \subset \mathcal{A}$ we denote $\mathcal{F}_*(\mathcal{C})$ the types such that $k_{ab} \neq 0 \Rightarrow (a,b) \in \mathcal{C}^2$.

2.2 Generating functions on types

Let $\mathbf{z}$ be a matrix of complex coefficients z_{ab} and let $\mathbf{k}$ a type, we denote

$$\mathbf{z}^{\mathbf{k}} = \prod_{(a,b)\in\mathcal{A}^2} z_{ab}^{k_{ab}} \tag{2}$$

As in [5] we introduce the generating functions of matrix variable, of the form $f(\mathbf{z}) = \sum_{\mathbf{k}\in K(\mathcal{A})} a_{\mathbf{k}}\mathbf{z}^{\mathbf{k}}$, where $a_{\mathbf{k}}$ is a sequence of indexed by types. In particular

$$B(\mathbf{z}) = \sum_{\mathbf{k}\in K(\mathcal{A})} B_{\mathbf{k}}\mathbf{z}^{\mathbf{k}} = \prod_{a\in\mathcal{A}} \frac{1}{1-\sum_{b\in\mathcal{A}} z_{ab}}\,. \tag{3}$$

Let $\mathcal{F}$ be the operator defined by $\mathcal{F}f(\mathbf{z}) = \sum_{\mathbf{k}\in\mathcal{F}_*} a_{\mathbf{k}}\mathbf{z}^{\mathbf{k}}$, *i.e.* the generating function restricted to balanced types. In particular we show in [5] that

$$\mathcal{F}B(\mathbf{z}) = \frac{1}{\det(\mathbf{I}-\mathbf{z})}\,. \tag{4}$$

If $\mathcal{C} \subset \mathcal{A}$ we define by abuse of notation $K(\mathcal{C})$ as the subset of $K(\mathcal{A})$ such that $k_{ab} \neq 0 \Rightarrow (a,b) \in \mathcal{C}^2$. Therefore $\mathcal{F}_*(\mathcal{C}) = \mathcal{F}_* \cap K(\mathcal{C})$. We denote $\mathcal{G}_\mathcal{C}$ the operator

$$\mathcal{G}_\mathcal{C} f(\mathbf{z}) = \sum_{\mathbf{k}\in K(\mathcal{C})} a_{\mathbf{k}}\mathbf{z}^{\mathbf{k}}\,. \tag{5}$$

The operators $\mathcal{F}$ and $\mathcal{G}_\mathcal{C}$ commute and we have

$$\mathcal{G}_\mathcal{C}\mathcal{F}B(\mathbf{z}) = \frac{1}{\det_\mathcal{C}(\mathbf{I}-\mathbf{z})} \tag{6}$$

where $\det_\mathcal{C}(\mathbf{M})$ for a matrix $\mathbf{M}$ defined on $\mathcal{A}\times\mathcal{A}$ is the determinant of the matrix $\mathbf{M}$ with only the row and columns indexes in $\mathcal{C}$.

2.3 Enumerating Eulerian paths

If X_n starts with a and ends with symbol b, and has type $\mathbf{k}$ then $\mathbf{k}+\delta_{ba} \in \mathcal{F}_n$ with δ_{ab} the type with all coefficients at zero except the (a,b) coefficient set at one. Let $E_{\mathbf{k}}$ the number of Eulerian circuit that can be made from a string of type $\mathbf{k}$ of the syntax graph. We use the following definitions: $\mathbf{k}^*$ is the matrix whose (a,b) coefficients is $\frac{k_{ab}}{k_a}$.

Theorem 1 *Let a type* $\mathbf{k}$ *such that* $\mathbf{k}+\delta_{ba} \in \mathcal{F}_n$ *(thus* X_n *ends with symbol* b*), we have*

$$E_{\mathbf{k}} = B_{\mathbf{k}+\delta_{ba}} \det_{\mathcal{A}-\{a\}}(\mathbf{I}-\mathbf{k}^*)(1+O(\frac{1}{\sqrt{n}})) \tag{7}$$

and

$$N^a_{\mathbf{k}} = B_{\mathbf{k}} \det_{\mathcal{A}-\{b\}}(\mathbf{I}-\mathbf{k}^*)(1+O(\frac{1}{\sqrt{n}})) \tag{8}$$

Most of the analysis is in [5] but for the sake of completeness we again give the proof of this theorem, because it leads to an important insight in the enumeration of types and Eulerian paths. We first concentrate on $E_{\mathbf{k}}$ estimate. We denote $F^a_{\mathbf{k}}$ the number of Eulerian circuit that can be made from a string starting from symbol a with type $\mathbf{k}$ *after* being made cyclic. In other words if $\mathbf{k} \neq 0$:

$$F^a_{\mathbf{k}} = \sum_{b \in \mathcal{A}} E_{\mathbf{k}-\delta_{ba}} \ . \tag{9}$$

Let $F^a(\mathbf{z})$ its generating function. We also denote $F^{ba}_{\mathbf{k}}$ the number of Eulerian paths that can be made starting from symbol a and ending on symbol b, or equivalently the number of Eulerian path starting with edge ba. We have:

$$F^{ba}_{\mathbf{k}} = N^a_{\mathbf{k}-\delta_{ba}} \ . \tag{10}$$

Lemma 1 *Let* $f(\mathbf{z})$ *and* $g(\mathbf{z})$ *be two generating functions. If* $\mathbf{k}$ *is a balanced type:*

$$[\mathbf{z}^{\mathbf{k}}]\mathcal{F}f(\mathbf{z})\mathcal{F}g(\mathbf{z}) = [\mathbf{z}^{\mathbf{k}}]f(\mathbf{z})\mathcal{F}g(\mathbf{z}) \ . \tag{11}$$

Proof: Since for all balanced type $\mathbf{k}$, $[\mathbf{z}^{\mathbf{k}}]f(\mathbf{z}) = [\mathbf{z}^{\mathbf{k}}]\mathcal{F}f(\mathbf{z})$, we have: $\mathcal{F}(f(\mathbf{z})\mathcal{F}g(\mathbf{z})) = \mathcal{F}f(\mathbf{z})\mathcal{F}g(\mathbf{z})$. □

Lemma 2 *We have the identity*

$$\left\{ \begin{array}{rcl} F^a(\mathbf{z}) & = & \frac{\det_{\mathcal{A}-\{a\}}(\mathbf{I}-\mathbf{z})}{\det(\mathbf{I}-\mathbf{z})} - 1 \ , \\ F^{ba}(\mathbf{z}) & = & \frac{\mathcal{F}(B(\mathbf{z})z_{ba})}{\det_{\mathcal{A}-\{b\}}(\mathbf{I}-\mathbf{z})} \ , \end{array} \right. \tag{12}$$

and

$$\left\{ \begin{array}{rcl} E_{\mathbf{k}} & = & [\mathbf{z}^{\mathbf{k}+\delta_{ba}}]B(\mathbf{z})\det_{\mathcal{A}-\{a\}}(\mathbf{I}-\mathbf{z}) \ , \\ N^a_{\mathbf{k}} & = & [z^{\mathbf{k}+\delta_{ba}}]B(\mathbf{z})z_{ba}\det_{\mathcal{A}-\{b\}}(\mathbf{I}-\mathbf{z}) \ . \end{array} \right. \tag{13}$$

Proof: Let $\mathbf{k}$ be a balanced type. If we take an arbitrary rotation among the $B_{\mathbf{k}}$ rotations and we proceed to create the path from symbol a: If $k_a = 0$ there is no path, and therefore $\mathbf{k} \in \mathcal{F}_*(\mathcal{A}-\{a\})$. Otherwise the path creates an anchored sub-Eulerian circuit (*i.e.*, an Eulerian circuit on a subgraph) that will end on symbol a after exhausting all the edges going to and departing from a. If $\mathbf{k}'$ is the type of this sub-Eulerian circuit then a rotation remains on $\mathbf{k}-\mathbf{k}'$ that does not contain symbol a, thus $\mathbf{k}-\mathbf{k}' \in \mathcal{F}_*(\mathcal{A}-\{a\})$. Therefore if $\mathbf{k} \notin \mathcal{F}_*(\mathcal{A}-\{a\})$:

$$B_{\mathbf{k}} = \sum_{\mathbf{k}-\mathbf{k}' \in \mathcal{F}_*(\mathcal{A}-\{a\})} F^a_{\mathbf{k}'} B_{\mathbf{k}-\mathbf{k}'} \ . \tag{14}$$

Thus we have the identity

$$\mathcal{F}B(\mathbf{z}) - \mathcal{G}_{\mathcal{A}-\{a\}}\mathcal{F}B(\mathbf{z}) = F^a(\mathbf{z})\mathcal{G}_{\mathcal{A}-\{a\}}\mathcal{F}B(\mathbf{z}) \; . \tag{15}$$

Concerning the $F^{ba}_{\mathbf{k}}$, we start the rotation from symbol b with the edge ba therefore there are $B_{\mathbf{k}-\delta_{ba}}$ such rotations, thus we have the identities

$$B_{k-\delta_{ba}} = \sum_{\mathbf{k}-\mathbf{k}' \in \mathcal{F}_*(\mathcal{A}-\{b\}} F^{ba}_{\mathbf{k}'} B_{\mathbf{k}-\mathbf{k}'} \; , \tag{16}$$

and

$$\mathcal{F}(B(\mathbf{z})z_{ba}) = F^{ba}(\mathbf{z})\mathcal{G}_{\mathcal{A}-\{b\}}\mathcal{F}B(\mathbf{z}) \; . \tag{17}$$

Thus

$$\left\{ \begin{array}{rcl} F^a_{\mathbf{k}} & = & [\mathbf{z}^{\mathbf{k}}]\mathcal{F}B(\mathbf{z})\text{det}_{\mathcal{A}-\{a\}}(\mathbf{I}-\mathbf{z}) \\ F^{ba}_{\mathbf{k}} & = & [\mathbf{z}^{\mathbf{k}}]\mathcal{F}(B(\mathbf{z})z_{ba})\text{det}_{\mathcal{A}-\{a\}}(\mathbf{I}-\mathbf{z}) \end{array} \right. \tag{18}$$

Since $\mathbf{k}$ a balanced type, using lemma 1 achieves the proof. □

The end of the proof proceeds by the following technical lemma:

Lemma 3 *Let a generating function of the form* $B(\mathbf{z})f(\mathbf{z}) = \sum_{\mathbf{k}\in K(\mathcal{A})} a_{\mathbf{k}}\mathbf{z}^{\mathbf{k}}$. *Let* $\mathbf{k} \in \mathcal{F}_n$ *assume that* $f(\mathbf{z})$ *has no singularities, then the* $\mathbf{k}$*'s coefficient of* $[\mathbf{z}^{\mathbf{k}}]B(\mathbf{z})f(\mathbf{z})$ *satisfies:*

$$[\mathbf{z}^{\mathbf{k}}]B(\mathbf{z})f(\mathbf{z}) = B_{\mathbf{k}}f(\mathbf{k}^*) + O(\frac{1}{n})B_{\mathbf{k}} \; . \tag{19}$$

Proof: This is a variation of Theorem 1 (iii) in [5] that we try to generalize with the methods in [6]. See Appendix. In fact a more careful proof would give an $O(\frac{1}{n})$ error term. □

Remark When $f(\mathbf{z})$ has singularities one must check via a similar analysis that the contribution of $\mathbf{z}^{-\mathbf{k}}$ on these singularities does not exceed the order $(\mathbf{k}^*)^{-\mathbf{k}}$.

Corollary 1 *The probability that a random rotation starting from symbol* a *gives an Eulerian circuit in a string of length* n *and type* $\mathbf{k}$ *is equal to* $\text{det}_{\mathcal{A}-\{a\}}(\mathbf{I}-\mathbf{k}^*) + O(\frac{1}{n})$.

In passing, when $\mathbf{k}$ is proportional to $\mathbf{1}$ this probability tends to $\frac{1}{V}$.

2.4 Enumeration Eulerian components

To evaluate the number of components we will assume that the symbols $\mathcal{A}$ are ordered: $\mathcal{A} = \{a_1, \ldots, a_V\}$. We denote $\mathcal{A}_i = \mathcal{A} - \{a_1, \ldots, a_i\}$ with the convention $\mathcal{A}_0 = \mathcal{A}$. We also have $\mathcal{A}_V = \emptyset$. A rotation on a type $\mathbf{k}$ is decomposed the following ways:

- if $k_{a_1} = 0$ then start the Eulerian circuit on a_2, *etc*, on the first a_i such that $k_{a_i} \neq 0$.
- the sub Eulerian circuit anchored at a_i is one component and has type $\mathbf{k}'$, the remaining type $\mathbf{k}-\mathbf{k}' \in \mathcal{F}_*(\mathcal{A}_i)$, if it is non null, we continue the decomposition on a_{i+1}, *etc*.

Let $\mathbf{k}$ a balanced type and let r be an integer (smaller than or equal to V). We denote $C_{\mathbf{k}}^r$ the number of rotations among the $B_{\mathbf{k}}$ rotations on $\mathbf{k}$ that generate exactly r components. We also denote the enumerating generating function $C_{\mathbf{k}}(u) = \sum_r u^r C_{\mathbf{k}}^r$ and finally $C(\mathbf{z}, u) = \sum_{\mathbf{k} \in \mathcal{F}_*} C_{\mathbf{k}}(u)$.

Theorem 2 *We have the identity*

$$C(\mathbf{z}, u) = \prod_{i=1}^{V} \left(1 + (\frac{\mathcal{G}_{\mathcal{A}_{i-1}} \mathcal{F}B(\mathbf{z})}{\mathcal{G}_{\mathcal{A}_i} \mathcal{F}B(\mathbf{z})} - 1)u \right) \tag{20}$$

with the convention $\mathcal{G}(\emptyset)B(\mathbf{z}) = 1$.

Proof: We basically show by recursion: if $\mathbf{k} \notin \mathcal{F}_*(\mathcal{A}_1)$ then we have

$$C_{\mathbf{k}}(u) = \sum_{\mathbf{k}-\mathbf{k}' \in \mathcal{F}_*(\mathcal{A}_1)} u F_{\mathbf{k}'}^{a_1} E_{\mathbf{k}-\mathbf{k}'}(u) \; . \tag{21}$$

Going to generating functions we got the equation

$$C(\mathbf{z}, u) = u F^{a_1}(\mathbf{z}) \mathcal{G}_{\mathcal{A}_1} C(\mathbf{z}, u) + \mathcal{G}_{\mathcal{A}_1} C(\mathbf{z}, u) \; , \tag{22}$$

and the result come by developing the induction (using the fact that $\mathcal{G}(\mathcal{A}_1)(C(\mathbf{z}, u)$ concerns types in $\mathcal{F}_*(\mathcal{A}_1)$ and rotations starting from symbol a_2. □

Theorem 3 *We have the identity*

$$C_{\mathbf{k}}(u) = B_{\mathbf{k}} \prod_{i=1}^{V} (1 + (u-1) d_{\mathbf{k}}^i) + O(\frac{1}{n}) B_{\mathbf{k}} \tag{23}$$

with

$$d_{\mathbf{k}}^i = 1 - \frac{\det_{\mathcal{A}_{i-1}}(\mathbf{I} - \mathbf{k}^*)}{\det_{\mathcal{A}_i}(\mathbf{I} - \mathbf{k}^*)}. \tag{24}$$

Remark We have $d_{\mathbf{k}}^1 \to 1$ since always $\det(\mathbf{I} - \mathbf{k}^*) = 0$ (the row sums to 0).

Remark When $\mathbf{k}$ is proportional to $\mathbf{1}$ we have $\mathbf{k}^* = \frac{1}{V}\mathbf{1}$ and $\det_{\mathcal{A}_i}(\mathbf{I} - \mathbf{k}^*) = \frac{i}{V}$, thus $d_{\mathbf{k}}^i = \frac{1}{i}$. When V is large $C_{\mathbf{k}}(u) = B_{\mathbf{k}} u \exp((u-1)H(V)) G(u)(1 + O(\frac{1}{V})) + O(\frac{1}{n}) B_{\mathbf{k}}$ where $G(u)$ is the p.g.f of a random variable independent of V. In other words the number of components tends to be the sum of a Poisson variable of mean $H(V) = \log V - \gamma + O(\frac{1}{V})$ plus a fixed random variable.

Proof: We have

$$C(\mathbf{z}, u) = \mathcal{F}B(\mathbf{z}) \prod_{i=1}^{V} \left(\frac{\mathcal{G}_{\mathcal{A}_i} \mathcal{F}B(\mathbf{z})}{\mathcal{G}_{\mathcal{A}_{i-1}} \mathcal{F}B(\mathbf{z})} (1-u) + u \right) \tag{25}$$

We identify $\mathcal{G}_{\mathcal{A}_i} \mathcal{F}B(\mathbf{z}) = (\det_{\mathcal{A}_i}(\mathbf{I} - \mathbf{z}))^{-1}$ thus

$$C_{\mathbf{k}}(u) = [\mathbf{z}^{\mathbf{k}}] B(\mathbf{z}) \prod_{i=1}^{V} \left(\frac{\det_{\mathcal{A}_{i-1}}(\mathbf{I} - \mathbf{z})}{\det_{\mathcal{A}_i}(\mathbf{I} - \mathbf{z})} (1-u) + u \right) \tag{26}$$

and we end the proof with the application of lemma 3 (but with some care about the singularities of $\frac{\det_{\mathcal{A}_{i-1}}(\mathbf{I}-\mathbf{z})}{\det_{\mathcal{A}_i}(\mathbf{I}-\mathbf{z})}$ that we don't detail here). □

Corollary 2 *The average value $C'_{\mathbf{k}}(1)$ satisfies:*

$$C'_{\mathbf{k}}(1) = B_{\mathbf{k}} \sum_{i=1}^{V} d^i_{\mathbf{k}} + O(\frac{1}{n})B_{\mathbf{k}} \, . \tag{27}$$

Remark When $\mathbf{k}$ is proportional to $\mathbf{1}$, $C'_{\mathbf{k}}(1) = B_{\mathbf{k}}H(V) + O(\frac{1}{n})B_{\mathbf{k}}$. When V is large the average number of components is equivalent to $\log V$.

3 Analysis of rotations and Eulerian paths in Markov models

In this section we assume that the string X_n is generated by a Markov model with transition matrix $\mathbf{P}$. To simplify we assume that all strings start with the same fixed symbol a. For the sequel of the paper the symbol a is fixed. If $\mathbf{k}$ is the type of string X_n then $P(X_n) = \mathbf{P}^{\mathbf{k}}$.

The number of strings that share the same type $\mathbf{k}$ is exactly $N^a_{\mathbf{k}}$, thus we have the identity

$$\sum_{b\in\mathcal{A}} \sum_{\mathbf{k}+\delta_{ba}\in\mathcal{F}_n} N^a_{\mathbf{k}}\mathbf{P}^{\mathbf{k}} = 1 \, , \tag{28}$$

the sums being split according to the ending symbol of the string, (assuming all strings start with symbol a).

3.1 Average logarithm of the number of rotations

We denote L_n the average logarithm of the number of rotations that can be done from a random Markov string X_n made cyclic. This number is important because it gives the quantity of information that can be carried by a rotation in a Markov string.

Theorem 4 *The average quantity of information carried by a rotation in a random Markov string X_n is asymptotically equivalent to the entropy $H(X_n)$ of the string.*

For $c \in \mathcal{A}$ we denote $k_c = \sum_{d\in\mathcal{A}} k_{cd}$and $k^c = \sum_{d\in\mathcal{A}} k_{dc}$ respectively the outdegree and indegree of symbol c in the syntaxic graph. Let $\mathcal{F}_n$ the set of *balanced* types, *i.e.* such that $\forall c \in \mathcal{A} : k_c = k^c$, and such $\sum_{(c,d)\in\mathcal{A}^2} k_{cd} = n$. It is clear that the definition were extended to non positive types then $\mathcal{F}_n$ would be a coset of the lattice made of types $\mathbf{k}$ such that $\forall c \in \mathcal{A}$: $k_c = k^c$. Therefore $\mathcal{F}_n$ is the intersection of this coset with the non negative type $\mathbf{k}$ such that $\sum_{(c,d)\in\mathcal{A}^2} k_{cd} = n$. We denote ω_V its elementary covolume (the volume between neighboring elements). In order to get into the proof we need several technical lemmas.

Lemma 4 *The set $\mathcal{F}_n$ is a lattice of dimension $V^2 - V$ [5] . Its size $a(n) = O(n^{V^2-V+1})$ and we denote ω the volume of its elementary element.*

Proof: The set of integer matrices is embedded in the vector space of real matrices which is a dimension V^2. The V balance equations are in fact $V - 1$ since any of them can be deduced from the sum of the other. There is a last equation to specify that all coefficients sum to n. □

We denote $\mathcal{F}(1)$ the set of balanced real matrices with positive coefficients that sum to 1. For $\mathbf{y}$ a real non negative matrix and $s \geq 0$, we denote $L(\mathbf{y}, s) = \sum_{(c,d)\in\mathcal{A}^2} y_{c,d} \log(\frac{y_c}{y_{cd}} p_{cd}^s)$. We denote $\mathbf{P}(s)$ the matrix made of coefficients $(p_{cd})^s$ (if $p_{cd} = 0$ we assume $(p_{cd})^s = 0$). We denote $\lambda(s)$ the main eigenvalues of matrix $\mathbf{P}(s)$.

Lemma 5 *Let $s > 0$, the maximal value of $L(\mathbf{y}, s)$ for $\mathbf{y} \in \mathcal{F}(1) - \frac{\delta_{ba}}{n}$ is the matrix $\tilde{\mathbf{y}}_n(s)$ which converges to a matrix $\tilde{\mathbf{y}}(s) \in \mathcal{F}(1)$ whose (c, d) coefficient is $v_c(s)u_d(s)\frac{p_{cd}^s}{\lambda(s)}$ with $(u_c(s))_{c\in\mathcal{A}}$ and $(v_c(s))_{c\in\mathcal{A}}$ being respectively the right and left main eigenvectors of $\mathbf{P}(s)$ (with $\sum_{c\in\mathcal{A}} u_c(s)v_c(s) = 1$). For instance $L(\tilde{\mathbf{y}}(s), s) = \log \lambda(s)$ and $L(\tilde{\mathbf{y}}_n(s), s) = L(\tilde{\mathbf{y}}(s), s) - \frac{1}{n}(\log \lambda(s)\frac{u_b(s)}{u_a(s)})$.*

Proof: See appendix. □

Proof of Theorem 4: For convenience we will handle only natural logarithms. We have $L_n = \sum_{b\in\mathcal{A}} L_n^b$ with

$$L_n^b = \sum_{\mathbf{k}+\delta_{ba}\in\mathcal{F}_n} N_{\mathbf{k}}^a \mathbf{P}^{\mathbf{k}} \log B_{\mathbf{k}+\delta_{ba}} = \sum_{\mathbf{k}+\delta_{ba}\in\mathcal{F}_n} \mathbf{P}^{\mathbf{k}} B_{\mathbf{k}} \det{}_{\mathcal{A}-\{a\}}(\mathbf{I} - \mathbf{k}^*) \log B_{\mathbf{k}+\delta_{ba}} (1 + O(\frac{1}{n})) \ . \quad (29)$$

We assume that the order $O(\frac{1}{n})$ is uniform but this is not a necessary assumption since we will only need to consider a shrinking neighborhood around the matrix $\tilde{\mathbf{y}}(1)$. Using the Stirling approximation: $k! = \sqrt{2\pi k} k^k e^{-k}(1 + O(\frac{1}{k}))$ and defining $\ell(\mathbf{y}) = \sum_{cd} y_{cd} \log \frac{y_c}{y_{cd}}$ we have

$$\begin{aligned} L_n^b &= (n + O(1)) \frac{1}{n^{(V-1)V/2}} \sum_{\mathbf{k}+\delta_{ba}\in\mathcal{F}_n} r_b(\mathbf{y}) \exp(nL(\mathbf{y}, 1))\ell(\mathbf{y}) \\ &\quad + O(\log n) \end{aligned} \quad (30)$$

where $r_b()$ is a rational function and $\mathbf{y}$ is the matrix of coefficients $y_{cd} = \frac{k_{cd}}{n}$.

According to Lemma 5 the maximum value of $L(\mathbf{y}, 1)$ is $\log \lambda(1) = 0$ and thus we already have $L_n^b = O(a(n)) = O(n^{V^2-V})$. Since $\tilde{\mathbf{y}(1)}$ is the maximum of $L(\mathbf{y}, 1)$ over $\mathcal{F}(1)$ there exists $A > 0$ such that $\forall \mathbf{y} \in \mathcal{F}(1)$: $L(\mathbf{y}, 1) \leq L(\tilde{\mathbf{y}}(1), 1) - A\|\mathbf{y} - \tilde{\mathbf{y}}(1)\|^2$, and when $\mathbf{y} + \frac{1}{n}\delta_{ba} \in \mathcal{F}(1)$

$$L(\mathbf{y}, 1) \leq L(\tilde{\mathbf{y}}_n(1)) - A\|\mathbf{y} - \tilde{\mathbf{y}}_n(1)\|^2 \quad (31)$$

Let $B > 0$: we define

$$L_n^b(B) = \frac{n}{n^{(V-1)V/2}} \sum_{\substack{\mathbf{k}+\delta_{ba}\in\mathcal{F}_n \\ |\mathbf{y}-\tilde{\mathbf{y}}_n| < B \log n/\sqrt{n}}} r_b(\mathbf{y}) \exp(nL(\mathbf{y}))\ell(\mathbf{y}) \quad (32)$$

From (31) we have $L_n^b = L_n^b(B)(1 + O(n^{-1}))$. Since function $L(\mathbf{y}, 1)$ is infinitely derivable and attains its maximum in $\mathcal{F}(1)$, which is zero, on $\tilde{\mathbf{y}}_n(1)$ we have for all $\mathbf{y} \in \mathcal{F}(1) - \frac{1}{n}\delta_{ba}$

$$L(\mathbf{y}, 1) = L(\tilde{\mathbf{y}}_n(1)) - D_n(\mathbf{y} - \tilde{\mathbf{y}}_n(1)) + O(\|\mathbf{y} - \tilde{\mathbf{y}}_n(1)\|^3) \ , \quad (33)$$

where $D_n()$ is a positive quadratic form obtained from the second derivative of $L(\mathbf{y}, 1)$ on $\tilde{\mathbf{y}}_n(1)$. The value of L_n will be attained in the vicinity of the maximum since $\exp(nL(\mathbf{y}, 1))$ behaves like a Dirac when

$n \to \infty$. Indeed we are in a kind of Saddle point application. Thus, since $L(\tilde{\mathbf{y}}_n(1)) = \frac{1}{n}\log\lambda(1)\frac{u_b(1)}{u_a(1)} = 0$ ($\lambda(1)=1$ and $\forall c \in \mathcal{A}$: $u_a(1)=1$)

$$\begin{aligned} L_n^b(B) &= \frac{1}{n^{(V-1)V/2}}(1+O(\frac{\log^3 n}{\sqrt{n}})) \\ &\sum_{\substack{\mathbf{k}+\delta_{ba}\in\mathcal{F}_n \\ |\mathbf{y}-\tilde{\mathbf{y}}_n|<B\log n/\sqrt{n}}} \tilde{r}(\mathbf{y}) e^{nD_n(\mathbf{y}-\tilde{\mathbf{y}}_n(1))} . \end{aligned} \tag{34}$$

Since $\frac{1}{n}\mathcal{F}_n$ is a lattice of degree $(V-1)V$ with elementary volume ωn^{-V^2+V} and $D_n()$ converge to some non negative quadratic form $D()$ on $\mathcal{F}(1)$:

$$\begin{aligned} L_n^b(B) &= n\ell(\tilde{\mathbf{y}}(1))\frac{r(\tilde{\mathbf{y}}(1))}{n^{(V-1)V/2}}\frac{n^{V^2-V}}{\omega}(1+O(\frac{1}{n})) \\ &\int_{\mathcal{F}(1)} e^{-nD(\mathbf{y}-\tilde{\mathbf{y}}(1))} d\mathbf{y}^{V(V-1)} \end{aligned} \tag{35}$$

We use the property that

$$\int_{\mathcal{F}(1)} \exp(-nD(\mathbf{y}-\tilde{\mathbf{y}}))d\mathbf{y}^{V(V-1)} = \frac{1}{\sqrt{n^{V^2-V}\det(\pi D)}} \tag{36}$$

where $\det(\pi D)$ must be understood as the determinant of the quadratic operator $\pi D()$ on the vector space $\mathcal{F}(1)$. Therefore

$$L_n = n\ell(\mathbf{y})\frac{1}{\sqrt{\det(\pi D)}}\sum_{b\in\mathcal{A}} r_b(\tilde{\mathbf{y}})(1+O(\frac{\log^3 n}{\sqrt{n}})) . \tag{37}$$

The same analysis can be done by removing $\log B_{\mathbf{k}+\delta_{ba}}$ and since via (28) we shall get $\sum_{b\in\mathcal{A}}\frac{r_b(\tilde{\mathbf{y}}(1))}{\sqrt{\det(\pi D)}} = 1$, we get

$$L_n = n\ell(\tilde{\mathbf{y}}(1))(1+O(\frac{\log^3 n}{\sqrt{n}})) . \tag{38}$$

We terminate the proof of theorem 4 by the fact that $n\ell(\tilde{\mathbf{y}}(1)) = H(X_n)$ (since $\forall c \in \mathcal{A}$: $u_c(1)=1$ and $(v_c(1))_{c\in\mathcal{A}}$ is the Markov stationary distribution). □

Remark The error term in $O(\frac{\log^3 n}{\sqrt{n}})$ is in fact very rough and could be arranged to be in $O(\frac{1}{n})$. We conjecture that $H(X_n) - L_n = O(\log n)$ as experimented in the last section.

3.2 Average number of rotations

Theorem 5 *The average number of rotations in a Markov string X_n of length n is equivalent to $\alpha\lambda^{2n}(\frac{1}{2})$ for some λ, and $\alpha > 0$ that can be explicitly computed.*

Proof: We proceed equivalently as in the proof of Theorem 4 except that $R_n = \sum_{b\in\mathcal{A}} R_n^b$ with

$$R_n^b = \sum_{\mathbf{k}+\delta_{ba}\in\mathcal{F}_n} N_{\mathbf{k}}^a \mathbf{P}^{\mathbf{k}} B_{\mathbf{k}+\delta_{ba}} = \sum_{\mathbf{k}+\delta_{ba}\in\mathcal{F}_n} \mathbf{P}^{\mathbf{k}} B_{\mathbf{k}}^2 \frac{k_b+1}{k_{ba}+1}\det\nolimits_{\mathcal{A}-\{a\}}(\mathbf{I}-\mathbf{k}^*)(1+O(\frac{1}{n})) . \tag{39}$$

The main factor $\mathbf{P}^{\mathbf{k}} B_{\mathbf{k}}^2$ leads to a factor $\exp(2nL(\mathbf{y}, \frac{1}{2}))$. Consideration on the order of the n factors leads to the estimate in $\alpha\lambda^{2n}(\frac{1}{2})$. □

Remark This average is purely theoretical, in practice it is hardly possible to calculate this result when n is large, since the most important and decisive contributions come from strings with extremely low probabilities.

3.3 Eulerian components

Theorem 6 *The probability p_n that a random rotation on a Markov string of length n builds a full Eulerian path (i.e. results in a single Eulerian component) tends to* $\det_{\mathcal{A}-\{a\}}(\mathbf{I}-\mathbf{P})$ *when $n \to \infty$. The p.g.f of the number of Eulerian components built from a random rotation $C_n(u)$ tends to*

$$\prod_{i=1}^{V} \left(\frac{\det_{\mathcal{A}_{i-1}}(\mathbf{I}-\mathbf{P})}{\det_{\mathcal{A}_i}(\mathbf{I}-\mathbf{P})}(1-u)+u \right) .$$

Remark Notice that the first factor is u since $\det(\mathbf{I}-\mathbf{P}) = 0$.

Proof: We have

$$p_n = \sum_{b\in\mathcal{A}} \sum_{\mathbf{k}+\delta_{ba}\in\mathcal{F}_n} N_{\mathbf{k}}^a \mathbf{P}^{\mathbf{k}} \det_{\mathcal{A}-\{a\}}(\mathbf{I}-\mathbf{k}^*)(1+O(\frac{1}{n})) . \tag{40}$$

Then the proof proceeds as with the proof of theorem 4 but replacing $\ell(\mathbf{y})$ with $\det_{\mathcal{A}-\{a\}}(\mathbf{I}-\mathbf{y}^*)$, thus we get

$$p_n \sim \det_{\mathcal{A}-\{a\}}(\mathbf{I}-\tilde{\mathbf{y}}(1)) , \tag{41}$$

and we get the result thanks to the identity $\tilde{\mathbf{y}}(1) = \mathbf{P}$. Similarly

$$C_n(u) = \sum_{b\in\mathcal{A}} \sum_{\mathbf{k}+\delta_{ba}} \mathbf{P}^{\mathbf{k}} N_{\mathbf{k}}^a \frac{C_{\mathbf{k}+\delta_{ba}}(u)}{B_{\mathbf{k}+\delta_{ba}}} \tag{42}$$

$$= \sum_{b\in\mathcal{A}} \sum_{\mathbf{k}+\delta_{ba}} \mathbf{P}^{\mathbf{k}} N_{\mathbf{k}}^a \prod_{i=1}^{V} \left(\frac{\det_{\mathcal{A}_{i-1}}(\mathbf{I}-(\mathbf{k}+\delta_{ba})^*)}{\det_{\mathcal{A}_i}(\mathbf{I}-(\mathbf{k}+\delta_{ba})^*)}(1-u)+u \right) (1+O(\frac{1}{n})) . \tag{43}$$

Since $(\mathbf{k}+\delta_{ba})^* = \mathbf{k}^* + O(\frac{1}{n})$ we get by replacing $\ell(\mathbf{y})$ by $\prod_{i=1}^{V} \left(\frac{\det_{\mathcal{A}_{i-1}}(\mathbf{I}-\mathbf{y}^*)}{\det_{\mathcal{A}_i}(\mathbf{I}-\mathbf{y}^*)}(1-u)+u \right)$ in the proof of Theorem 4, to get our result. □

4 Experimental results and Conclusion

Figure 1 shows $I(X_n, Y_n)$, the discrepancy between L_n and $H(X_n)$ versus the string length n, when X_n is generated by a Markov process of memory 1 based on the statistics of the phrase "to be or not to be that is the question". As predicted the conjectured is quite sub-linear and seems to be in $\log n$ as assumed. Each point has been simulated 100 times.

In this paper, we evaluated the number of Eulerian circuits that could be obtained by an arbitrary rotation in a Markovian string given to a Markovian type. We also studied the number of Eulerian components that result from random rotations in Markovian strings. The asymptotic behaviour of those quantities was considered, when the size n of a string X tends to infinity.

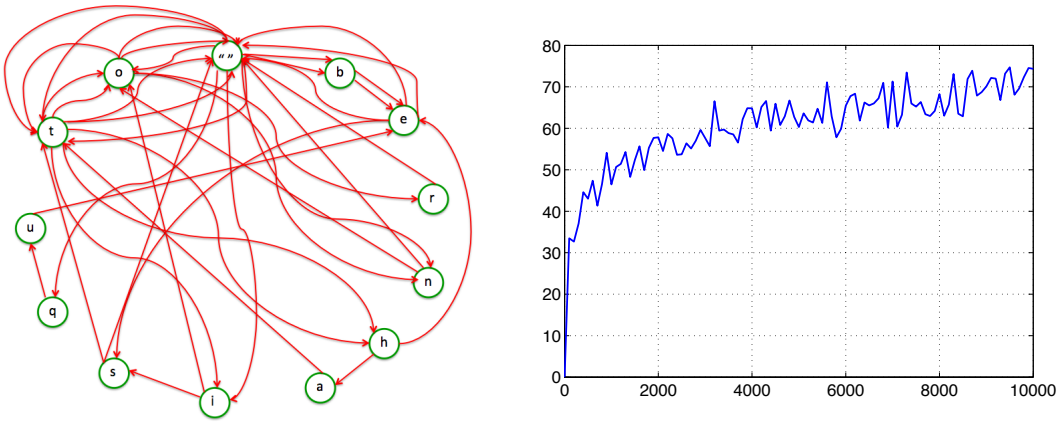

Fig. 1: Left: type of "*to be or not to be that is the question*", Right: Difference $H(X_n) - L_n$ (in bits) versus n.

References

[1] W. T. Tutte, C. A. B. Smith, *On unicursal paths in a network of degree 4*, American Mathematical Monthly, Vol. 48, pp. 233–237, 1941, JSTOR 2302716.1941.

[2] T. van Aardenne-Ehrenfest, N. G. de Bruijn, *Circuits and trees in oriented linear graphs"*, Simon Stevin Vol. 28, pp. 203–217, 1951.

[3] B. McKay and R. W. Robinson, *Asymptotic enumeration of Eulerian circuits in the complete graph*, Combinatorica, no. 4, 10, pp. 367–377, 1995.

[4] M. I. Isaev, *Asymptotic number of Eulerian circuits in complete bipartite graphs*, in Proc. of 52nd MFTI Conference, Moscow, 2009.

[5] P. Jacquet and W. Szpankowski, *Markov Types and Minimax Redundancy for Markov Sources*, in IEEE Transactions on Information Theory, Vol. 50, No. 7, July 2004.

[6] Pemantle, R., Wilson, M. C. *Asymptotics of multivariate sequences: I. smooth points of the singular variety* Journal of Combinatorial Theory, Series A, 97(1), 129-161. (2002).

Appendix

Proof of lemma 3: We give only a sketch. We proceed via multivariable cauchy:

$$a_{\mathbf{k}} = \left(\frac{1}{2i\pi}\right)^{V^2} \oint \dots \oint B(\mathbf{z}) f(\mathbf{z}) \frac{d^{V^2}\mathbf{z}}{\mathbf{z^{k+1}}} \,. \tag{44}$$

To simplify we assume that $f(\mathbf{z})$ has no singularities. We apply theorem 3.9 of [6], the Kernel $\mathcal{V}$ is the set of matrices $\mathbf{z}$ such that $\forall a \in \mathcal{A}$: $\sum_{b\in\mathcal{A}} z_{ab} = 1$, it is of dimension $V^2 - V$. The function $B(\mathbf{z})$ satisfies all the conditions of theorem 3.9 [6]. Thus

$$a_{\mathbf{k}} = (\mathbf{z}^*)^{-\mathbf{k}} f(\mathbf{z}^*) A(\mathbf{z}^*) n^{\frac{V-V^2}{2}} + O((\mathbf{z}^*)^{-\mathbf{k}} n^{\frac{V-V^2}{2}-1})) \,, \tag{45}$$

where $\mathbf{z}^*$ is the element of $\mathcal{V}$ that attains the maximal value of $\mathbf{z}^{\mathbf{k}}$ and $A(.)$ is a specific function detailed in [6]. The maximal value of $\mathbf{z}^{\mathbf{k}}$ on $\mathcal{V}$ satisfies for all $a \in \mathcal{A}^2$ the quantities $\frac{\partial}{\partial z_{ab}} \log \mathbf{z}^{\mathbf{k}}$ must be identical. Since

$$\frac{\partial}{\partial z_{ab}} \log \mathbf{z}^{\mathbf{k}} = \frac{k_{ab}}{z_{ab}} \,, \tag{46}$$

we have $\mathbf{z}^* = \mathbf{k}^*$. Since the estimate of $B_{\mathbf{k}}$ is given by replacing $f(\mathbf{z})$ by 1:

$$B_{\mathbf{k}} = (\mathbf{z}^*)^{-\mathbf{k}} A(\mathbf{z}^*) n^{\frac{V-V^2}{2}} + O((\mathbf{z}^*)^{-\mathbf{k}} n^{\frac{V-V^2}{2}-1})) \,, \tag{47}$$

We get the claimed result. □

Remark When $f(\mathbf{z})$ has singularities one must check via a similar analysis that the contribution of $\mathbf{z}^{-\mathbf{k}}$ does not exceed the order $(\mathbf{k}^*)^{-\mathbf{k}}$.

Proof of lemma 5: We have for all $(c,d) \in \mathcal{A}^2$ the gradient matrix is

$$\frac{\partial}{\partial y_{cd}} L(\mathbf{y}, s) = \log \frac{y_c}{y_{cd}} p_{cd}^s \,. \tag{48}$$

The maximum on $\mathcal{F}(1)$ or $\mathcal{F}(1) - \frac{1}{n}\delta_{ba}$ must be member of the vector space generated by the matrix $\mathbf{1}$ (made of all one), and the matrices $\mathbf{A}_j$, $j \in \mathcal{A}$, the coefficients of A_j are all zeros excepted 1 on the jth column and -1 on the jth row, and zero on the diagonal. These matrices are the orthogonal matrices that define $\mathcal{F}(1)$ (or a translation of it). Membership to this vector space is equivalent to the fact that $\frac{\partial}{\partial y_{cd}} L(\mathbf{y}, s)$ must be of the form $\alpha + z_c - z_d$ for some α and $(z_c)_{c\in\mathcal{A}}$, which is equivalent to the fact that

$$\frac{y_{cd}}{y_c} = \frac{x_d}{x_c} \frac{p_{cd}^s}{\lambda} \tag{49}$$

for some λ and $(x_c)_{c\in\mathcal{A}}$. From the fact that $\sum_{d\in\mathcal{A}} \frac{y_{cd}}{y_c} = 1$ we get $\lambda = \lambda(s)$ and $(x_c)_{c\in\mathcal{A}} = (u_c(s))_{c\in\mathcal{A}}$, *i.e.* $\lambda x_c = \sum_{d\in\mathcal{A}} p_{cd}^s x_d$. Consequently

$$\begin{aligned} L(\mathbf{y}, s) &= \sum_{(c,d)\in\mathcal{A}^2} y_{cd} \log(\lambda \frac{x_c}{x_d}) && (50)\\ &= \log(\lambda) \sum_{(c,d)\in\mathcal{A}^2} y_{cd} && (51)\\ &\quad + \sum_{c\in\mathcal{A}} (y_c - y^c) \log(x_c) && (52) \end{aligned}$$

Thus $L(\tilde{\mathbf{y}}(s), s) = \log \lambda(s)$ and $L(\tilde{\mathbf{y}}_n(s), s) = (1 - \frac{1}{n}) \log \lambda(s) + \frac{1}{n} \log \frac{u_a(s)}{u_b(s)}$. □