A PROBABILISTIC ANALYSIS OF A PATTERN MATCHING PROBLEM

July 9, 1992

Mikhail J. Atallah*
Dept. of Computer Science
Purdue University
W. Lafayette, IN 47907
U.S.A.

Philippe Jacquet†
INRIA
Rocquencourt
78153 Le Chesnay Cedex
France

Wojciech Szpankowski‡
Dept. of Computer Science
Purdue University
W. Lafayette, IN 47907
U.S.A.

Abstract

The study and comparison of strings of symbols from a finite or an infinite alphabet is relevant to various areas of science, notably molecular biology, speech recognition, and computer science. In particular, the problem of finding the minimum "distance" between two strings (in general, two blocks of data) is of great importance. In this paper we investigate the (string) pattern matching problem in a probabilistic framework, namely, it is assumed that both strings form an independent sequences of i.i.d. symbols. Given a text string **a** of length n and a pattern string **b** of length m, let $M_{m,n}$ be the maximum number of matches between **b** and all m-substrings of **a**. Our main probabilistic result shows that $M_{m,n} - mP = \Theta(\sqrt{2m \log n})$ *in probability* (pr.) provided $m, n \to \infty$ such that $\log n = o(m)$, where P is the probability of a match between any two symbols of these strings. If the alphabet is symmetric (i.e., all symbols occur with the same probability), a stronger result holds, namely $(M_{m,n} - mP) \sim \sqrt{2m(1/V - 1/V^2) \log n}$ (pr.), where V is the size of the alphabet. In either case, symmetric alphabet or not, we also prove that $M_{m,n}/m \to P$ *almost surely* (a.s).

*This author's research was supported by the Office of Naval Research under Grants N0014-84-K-0502 and N0014-86-K-0689, and in part by AFOSR Grant 90-0107, and the NSF under Grant DCR-8451393, and in part by Grant R01 LM05118 from the National Library of Medicine.

†This research was primary supported by NATO Collaborative Grant 0057/89.

‡This author's research was supported by AFOSR Grant 90-0107 and NATO Collaborative Grant 0057/89, and, in part by the NSF Grant CCR-8900305, and by Grant R01 LM05118 from the National Library of Medicine.

1. INTRODUCTION

Pattern matching is one of the most fundamental problems in computer science. It is relevant to various areas of science, notably molecular biology, speech recognition, theory of languages, coding, and so forth. For example, in molecular biology often one compares two, say DNA, sequences to detect an unusual homology (similarity) between these strings. Comparisons of blocks of data representing, say time series of physical measurements, prices, etc., also lead to string comparisons, but with possibly infinite alphabets. In general, the problem of finding the minimum "distance" between two strings is of great importance.

The version of the string matching problem we investigate here is the following one. Consider two strings, a *text* string $\mathbf{a} = a_1 a_2 ... a_n$ and a *pattern* string $\mathbf{b} = b_1 b_2 ... b_m$ of lengths n and m respectively, such that symbols a_i and b_j belong to a V-ary alphabet $\Sigma = \{1, 2, ..., V\}$. The alphabet may be finite or not. Let C_i be the number of positions at which the substring $a_i a_{i+1} ... a_{i+m-1}$ agrees with the pattern $\mathbf{b}$ (the index j that is out of range is understood to stand for $1 + (j \bmod n)$). That is, $C_i = \sum_{j=1}^{m} equal(a_{i+j-1}, b_j)$ where $equal(x, y)$ is one if $x = y$, and zero otherwise. In other words, the "distance" between $\mathbf{a}$ and $\mathbf{b}$ is the Hamming distance. We are interested in the quantity $M_{m,n} = \max_{1 \le i \le n}\{C_i\}$ which represents the best matching between $\mathbf{b}$ and any m-substring of $\mathbf{a}$.

We analyze $M_{m,n}$ under the following probabilistic assumption: *symbols from the alphabet Σ are generated independently, and symbol $i \in \Sigma$ occurs with probability p_i*. This probabilistic model is known as the *Bernoulli model* [26]. An estimate of $M_{m,n}$ can be used in various statistical tests, for example in the usual hypothesis-testing context in which one needs to detect a "significant" matching. In that context, a matching is considered significant when it is much larger than the one obtained for *random* strings, that is, than $M_{m,n}$; indeed, if even random independent strings give rise to a match of $M_{m,n}$, it is reasonable not to attach particular significance to this amount of "similarity" between two strings. Furthermore, an estimate of $M_{m,n}$ can be used to control some parameters of algorithms designed for pattern matching, as we did in Atallah *et al.* [8].

Our main probabilistic result concerning $M_{m,n}$ can be summarized as follows. We prove that with high probability $M_{m,n} - mP = \Theta(\sqrt{2m \log n})$ for $n, m \to \infty$ such that $\log n = o(m)$ where P is the probability that two symbols match. The restriction on n and m is important. We explicitly give the constants that are implicit in the $\Theta(\cdot)$ notation. Our simulation experiments indicate that the second term in $M_{m,n}$, that is, $\Theta(\sqrt{2m \log n})$, spreads in probability between these two bounds, that is, we conjecture that the second term is rather a random variable than a constant. However, for the symmetric alphabet (i.e., all symbols occur with the same probability) stronger result holds, namely $(M_{m,n} - mP) \sim \sqrt{2m(1/V - 1/V^2)\log n}$ (pr.). In either case – symmetric alphabet or not – we demonstrate that *almost surely* (a.s.) $M_{m,n}/m \to P$ as long as $\log n = o(m)$.

A probabilistic analysis of any pattern matching is a "very complicated" problem as asserted by Arratia, Gordon and Waterman [7]. In fact, the literature on probabilistic analysis of pattern matching is very scanty. To the best of our knowledge, it is restricted to three papers of Arratia, Gordon and Waterman [5, 6, 7], however, only [7] is relevant to our study. Papers [5, 6] investigate the Erdös-Rényi law for the longest *contiguous* run of matches. Only in [7] is the number of matches in a *segment* of length m investigated. This last problem resembles the one discussed in this paper, however the authors of [7] restrict themselves to $\log n = \Theta(m)$ while we analyze the case when $\log n = o(m)$. Finally, it should be mentioned that our formulation of the problem is an attempt to study a more important problem of deletion, insertion and substitution developed during the evolution of nucleotide sequences (cf. Apostolico *et al.* [4], Louchard and Szpankowski [23]).

Algorithmic issues of pattern matching (with possible mismatches) are discussed in several papers. The problem of pattern matching with mismatches was posed by Galil [15] for the case of more than one mismatch. Of course, it has a linear time solution for the case of zero mismatch (i.e., $k = m$). For the case of a single mismatch (i.e., $k = m - 1$) a linear time solution is also known (attributed to Vishkin in [15]). The best known time bound for the general case of arbitrary k is $O(n\sqrt{m}\text{polylog}(m))$ and is due to Abrahamson [1]. The elegant probabilistic method of Chung and Lawler [9] led to an algorithm of $O((n/m)k \log m)$ time complexity on the average, assuming that $k < m/\log m + O(1)$ and that the alphabet size is $O(1)$.

The paper is organized as follows. The next section makes a more precise statement of our results. In particular, Theorem 2 contains our main probabilistic result. All proofs are delayed until Section 3, which is also of independent interest. It discusses a fairly general approach that can be used to analyze pattern matching in strings. In that section we apply extensively the saddle point method [18] to evaluate necessary asymptotic approximations.

2. MAIN RESULTS

This section presents our main results derived under the Bernoulli model discussed above. Note that, in such a model, $P = \sum_{i=1}^{V} p_i^2$ represents the probability of a match between a *given* position of the text string $\mathbf{a}$ and a given one in the pattern string $\mathbf{b}$. All our asymptotic results are valid in the so called *right domain asymptotic*, that is, for

$\log n = o(m)$. In fact, the most interesting range of values for m and n is such that $m = \Theta(n^\alpha)$ for some $0 \le \alpha \le 1$. However, to cover the whole right domain asymptotics, we assume that α may also slowly converge to zero, that is, $1 \le \alpha^{-1} \le o(m/\log m)$ (for simplicity of notation we write α instead of α_n).

Finally, for simplicity of the presentation, it helps to imagine that $\mathbf{a}$ is written on a cycle of size $n \ge 2m$. Then, we call $\rho^i(\mathbf{b})$ the version of $\mathbf{b}$ written on the same cycle, and cyclically shifted by i positions relative to $\mathbf{a}$.

Note that C_i is the number of matches between $\mathbf{a}$ and $\mathbf{b}$ when $\mathbf{b}$ is shifted by i positions. In our cyclic representation, C_i can alternatively be thought of as the number of places on the cycle in which $\mathbf{a}$ and $\rho^i(\mathbf{b})$ agree. It is easy to see that the distribution of C_i is binomial, hence for any i

$$\Pr\{C_i = \ell\} = \binom{m}{\ell} P^\ell (1-P)^{m-\ell} \ . \tag{1}$$

Naturally, the average number of matches EC_1 is equal to mP. Furthermore, C_i tends *almost surely* to its mean mP (by the *Strong Law of Large Numbers* [12]).

Now we can present our first result regarding $M_{m,n} = \max_{1\le i\le n}\{C_i\}$ which can be interpreted as the amount of similarity between $\mathbf{a}$ and $\mathbf{b}$.

Theorem 1 *If m and n are in the right domain asymptotic, that is , $\log n = o(m)$, then*

$$\lim_{m\to\infty} \frac{M_{m,n}}{m} = P \qquad \text{(a.s.)} \tag{2}$$

Proof. A lower bound on $M_{m,n}$ follows from the fact that, by its definition, $M_{m,n}$ must be greater than C_1 which tends – by the *Strong Law of Large Numbers* – almost surely to mP. So, now we concentrate on an upper bound. We first establish the bound *in probability* (pr.) sense. From Boole's inequality we have

$$\Pr\{M_{m,n} > r\} = \Pr\{C_1 > r \ \ or \ \ C_2 > r \ \ or \ \ \dots\ C_n > r\} \le n\Pr\{C_1 > r\} \ . \tag{3}$$

It suffices to show that for $r \sim mP$ the above probability becomes $o(1)$, that is, $n\Pr\{C_1 > (1+\epsilon)mP\} = o(1)$. For this we need an estimate of the tail for the binomial distribution (1). Such an estimate is computed in Section 2 by the saddle point method. A simpler (and more general) approach, however, is necessary for the purpose of this proof. We note that C_1 can be represented as a sum of m independent Bernoulli distributed random variables X_i, where X_i is equal to one when there is a match at the ith position, and zero otherwise. From the *Central Limit Theorem* we know that $(C_1 - EC_1)/(\sqrt{mP(1-P)}) \to$

$\mathcal{N}(0,1)$, where $\mathcal{N}(0,1)$ is the standard normal distribution. Let $G_m(x)$ be the distribution of $(\sum_{i=1}^m X_i - mEX_1)/\sqrt{m \cdot varX_1}$, and let $\Phi(x)$ be the standard normal distribution. Then, from Feller [12]

$$G_m(x) = \Phi(x) + \frac{e^{-x^2/2}}{\sqrt{2\pi}} o(\sqrt{m}) \ ,$$

where $\Phi(x) \sim \frac{e^{-x^2/2}}{x\sqrt{2\pi}}$. In (3), set $r = mP + (1+\varepsilon)\sqrt{m2P(1-P)\log n}$. Then

$$\Pr\{M_{m,n} > mP + (1+\varepsilon)\sqrt{2mP(1-P)\log n}\} \le \frac{1}{n^\varepsilon} \ . \tag{4}$$

The above rate of convergence does not yet warrant the application of the Borel-Cantelli lemma to prove (2), but clearly (4) proves $M_{m,n}/m \to P$ (pr.) in the right domain asymptotic.

To establish the stronger (a.s.) convergence, we proceed as follows. Define a subsequence $n_r = s2^r$ for some integers r and s. Let $\widetilde{M}_n = (M_{m,n} - mP)/\sqrt{2m}$. Then, by (4) $\widetilde{M}_{n_r}/\sqrt{\log n_r} \le \sqrt{P - P^2}$ (a.s.). To finish the proof we need to extend this to all n. Note that for a fixed s we can always find an r such that $s2^r \le n \le (s+1)2^r$. Furthermore, since $\widetilde{M}_n$ is a *stochastically increasing* sequence, we have

$$\lim_{n\to\infty} \sup \frac{\widetilde{M}_n}{\sqrt{\log n}} \le \lim_{r\to\infty} \sup \frac{\widetilde{M}_{(s+1)2^r}}{\sqrt{\log(s+1)2^r}} \frac{\sqrt{\log(s+1)2^r}}{\sqrt{\log s2^r}} = \sqrt{P-P^2} \qquad (a.s.) \ .$$

This proves (2). ■

Theorem 1 does not provide much useful information, and an estimate of $M_{m,n}$ based on it would be a very poor one. From the proof of Theorem 1 we learn, however, that $M_{m,n} - EC_1 = O(\sqrt{m\log n})$, hence the next term in the asymptotics of $M_{m,n}$ plays a very significant role, and definitely cannot be omitted in any reasonable computation. The next theorem – our main result – provides an extension of Theorem 1, and shows how much the maximum $M_{m,n}$ differs from the average EC_1. The proof of the above theorem will be given in Section 3.

Theorem 2 *Under the hypothesis of Theorem 1 we also have for every* $\varepsilon > 0$

$$\lim_{n\to\infty} \Pr\{\underline{\delta}(1-\varepsilon) \le \frac{M_{m,n} - mP}{\sqrt{2m\log n}} < \bar{\delta}(1+\varepsilon)\} = 1 \ , \tag{5}$$

where

$$\underline{\delta} = \max\{\sqrt{(1-\alpha)(P-T)}, \ \sqrt{\min\{(P-T), \ (1-\alpha)\delta_1\}}\} \tag{6}$$

$$\bar{\delta} = \min\{\sqrt{P-P^2}, \sum_{j=1}^{V} p_j\sqrt{1-p_j}\} \ . \tag{7}$$

In the above,

$$\delta_1 = \frac{(P-T)(P-3P^2+2T)}{6(T-P^2)} \ ,$$

where $T = \sum_{i=1}^V p_i^3$. ∎

Remark 1. The lower bound (6) and the upper bound (7) are composed of two different expressions. None of the expressions is better than the other for the entire range of the probabilities $0 < p_i < 1$. With respect to the lower bound, we can show that there exist values of the probabilities $\{p_i\}_{i=1}^V$ such that either $\underline{\delta}^2 = (1-\alpha)(P-T)$ or $\underline{\delta}^2 = \min\{(P-T), (1-\alpha)\delta_1\}$. The latter case clearly can occur. Surprisingly enough, the former case may happen too. Indeed, let one of the probability p_i dominates all the others, e.g., $p_1 = 1-\varepsilon$. Then, we have either $\underline{\delta}^2 = (1-\alpha)\varepsilon + O(\varepsilon^2)$ or $\underline{\delta}^2 = \min\{2/3 \cdot (1-\alpha)\varepsilon, \ \varepsilon\} + O(\varepsilon^2)$. For α close to one the former bound is tighter. With respect to the upper bound, we have a similar situation. In most cases, $\bar{\delta}^2 = P - P^2$. However, sometimes $\bar{\delta} = \sum_{i=1}^V p_i\sqrt{1-p_i}$. Indeed, this occurs for example for a binary alphabet with $p_1 < 0.1$. □

Note that, when all the $p_i \to 1/V$ (the symmetric case) we have $\delta_1 \to \infty$. Thus, for all $\alpha < 1$ we have $\underline{\delta} = \bar{\delta} = \sqrt{1/V - 1/V^2}$. Thus, we obtain the following corollary.

Corollary 1 *If the alphabet is symmetric and* $\alpha < 1$, *then for every* $\varepsilon > 0$ *the following holds*

$$\lim_{n\to\infty} \Pr\{1-\varepsilon < \frac{M_{m,n} - mP}{\sqrt{2m(1/V - 1/V^2)\log n}} < 1+\varepsilon\} = 1 \ , \tag{8}$$

that is, $M_{m,n} - mP \sim \sqrt{2m(1/V-1/V^2)\log n}$ *(pr.).* ∎

In the last subsection we will extend the condition of the corollary to the limiting case $\alpha = 1$.

In order to verify our bounds for $M_{m,n}$ we have performed some simulation experiments, which are presented in Figures 1, 2, and 3. In these figures we plot the experimental distribution function $\Pr\{(M_{m,n} - mP)/\sqrt{mP(1-P)} \le x\}$ versus x. Our simulation experiments presented in Figure 1 confirm our theoretical result of Corollary 1. Namely, for the symmetric alphabet, $(M_{m,n} - mP)/\sqrt{m \log n}$ converges to a constant. But – surprisingly – this is not true for an asymmetric alphabet, as indicated by Figures 2 and 3. In fact, based on these simulations we conjecture, that $(M_{m,n} - mP)/\sqrt{m\log n}$ converges *in probability* to a random variable, say Z, which is *not* a degenerate one, except for the symmetric alphabet. A study of the limiting distribution of Z seems to be very difficult. Our Theorem 2 provides some information regarding the behavior of the tail of the distribution of Z. In particular,

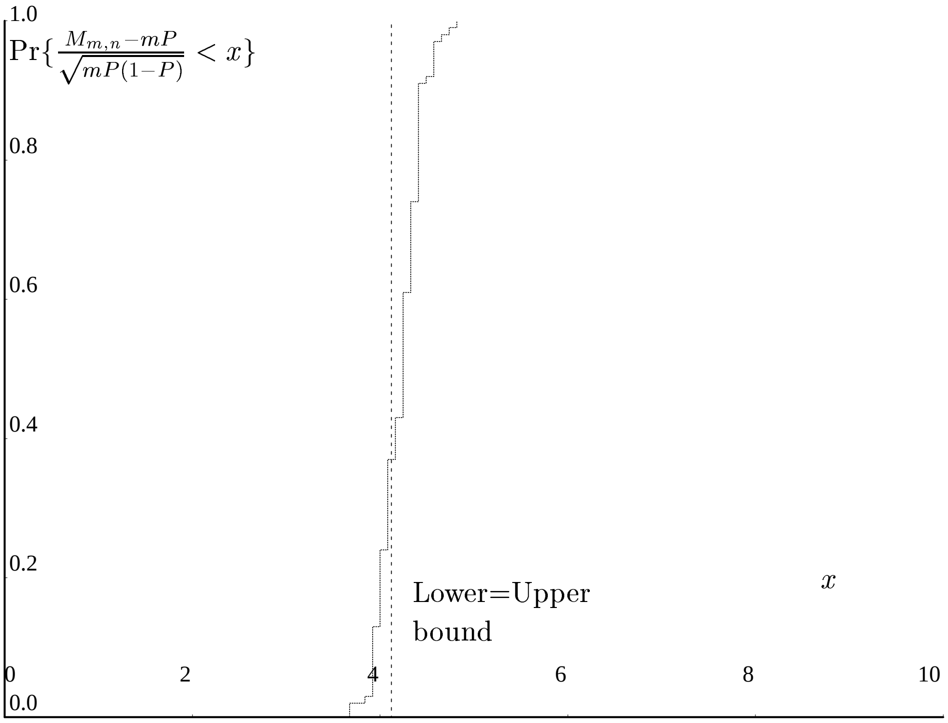

Figure 1: Distribution function of $(M_{m,n} - mP)(mP(1-P))^{-1/2}$ for $V = 2$ with $p_1 = 0.5$ (symmetric case), *via* simulations.

we know that Z degenerates for the symmetric alphabet, that is, possibly the variance of Z becomes asymptotically zero in this case.

3. ANALYSIS

In this section, we prove our main result, namely Theorem 2. In the course of deriving it, we establish some interesting combinatorial properties of pattern matching that have some similarities with the work of Guibas and Odlyzko [16, 17] (see also [19]). The proof itself consists of two parts: the upper bound (cf. Section 3.2) and the more difficult lower bound (cf. Section 3.3). Section 3.1 provides necessary mathematical background, and presents some preliminary results.

Before we plunge into technicalities, we present an overview of our approach. An upper bound on $M_{m,n}$ is obtained from (3), or a modification of it, and the real challenge is in establishing a tight lower bound. The idea was to apply the *second moment method*, in the form of Chung and Erdös [10] which states that for events $\{C_i > r\}$, the following holds

$$\Pr\{M_{m,n} > r\} = \Pr\{\bigcup_{i=1}^{n}(C_i > r)\} \geq \frac{(\sum_i \Pr\{C_i > r\})^2}{\sum_i \Pr\{C_i > r\} + \sum_{(i \neq j)} \Pr\{C_i > r \;\&\; C_j > r\}} . \quad (9)$$

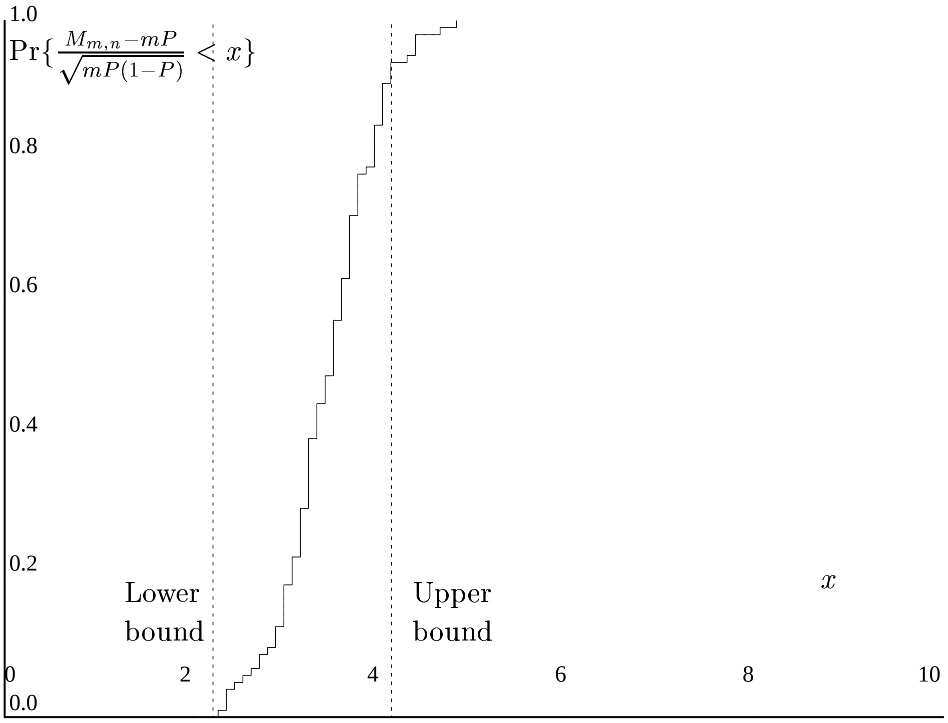

Figure 2: Distribution function of $(M_{m,n} - mP)(mP(1-P))^{-1/2}$ for $V = 2$ when $n = 50,000$ and $m = 600$, when $n = 50,000$ and $m = 600$, with $p_1 = 0.2$, *via* simulations.

Thus, one needs to estimate the joint distribution $\Pr\{C_i > r \;\&\; C_j > r\}$. Ideally, we would expect that the right-hand side of (9) goes to one when $r = mP + (1-\varepsilon)\sqrt{m2(P-P^2)\log n}$ for any $\varepsilon > 0$, which would match the upper bound.

Provided the above works, one needs an estimate of $\Pr\{C_i > r \;\&\; C_j > r\}$. This is a difficult task since C_1 and C_j are *strictly* positively correlated random variables for *all* values of $1 < j \leq n$. In fact, there are two types of dependency. For $j > m$, the patterns aligned at positions 1 and j do not overlap, but nevertheless C_1 and C_j are correlated since the same pattern is aligned. Although this correlation is not difficult to estimate, it is strong enough to "spoil" the second moment method. In addition, an even stronger correlation between C_1 and C_j occurs for $j < m$ since in addition overlapping takes place. We treat this type of dependency in Section 3.4.

To resurrect the second moment method, we introduce a *conditional* second moment method. The idea is to fix a string σ of length m built over the same alphabet as the pattern $\mathbf{b}$, and to estimate C_1 and C_j under the condition $\mathbf{b} = \sigma$. Then, C_1 and C_j are *conditionally independent* for all $j > m$, and the dependency occurs only for $j \leq m$, that is, when C_1 and C_j overlap. This idea we shall develop below.

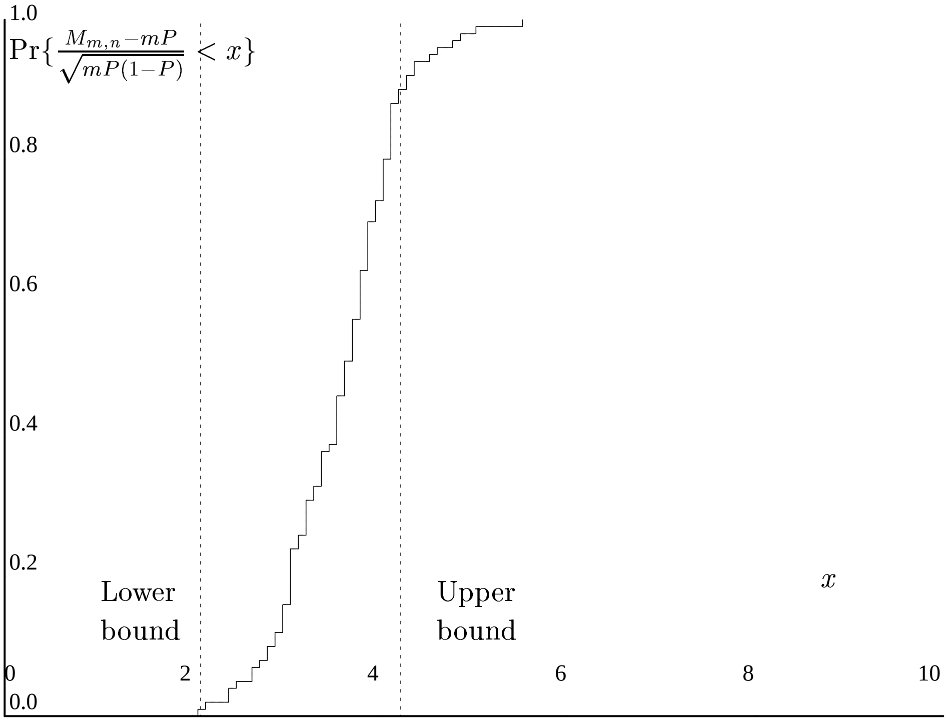

Figure 3: Distribution function of $(M_{m,n} - mP)(mP(1-P))^{-1/2}$ for $V = 2$ when $n = 100,000$ and $m = 1000$, with $p_1 = 0.1$, *via* simulations.

We introduce some notation. For a given σ, we define

$$G(r,\sigma) = \Pr\{C_1 > r|\mathbf{b} = \sigma\} \ , \tag{10}$$

$$F(r,\sigma) = \sum_{i=2}^{m} \Pr\{C_1 > r \ \& \ C_i > r \ |\mathbf{b} = \sigma\} \tag{11}$$

$$p(\sigma) = \Pr\{\mathbf{b} = \sigma\} \ . \tag{12}$$

Considering our Bernoulli model, $p(\sigma) = \prod_{j=1}^{j=V} p_j^{\Omega_j}$ where the Ω_j denotes the number of occurrence of the symbol j in the string σ. When σ varies, we may consider the conditional probabilities $G(r,\sigma)$ and $F(r,\sigma)$ as random variables with respect to σ. We adopt this point of view.

3.1 Preliminary Results

In this subsection, we put together several results concerning the tail of a binomially distributed random variable, that are used in the proof of our main results. If X is a binomially distributed random variable with parameters m and p, then we write $B(m,p,r) = \Pr\{X > r\}$ for the tail of X. We need the following result concerning the tail of C_1.

Lemma 1 *When m and τ both tend to infinity such that $\tau = O(\log m)$, then*

$$B(m, P, mP+\sqrt{m\tau}) = \Pr\{C_1 \geq mP + \sqrt{m\tau}\} \sim \frac{1}{\sqrt{2\pi(P-P^2)\tau}} \exp\left(\frac{-\tau}{2(P-P^2)}\right) . \quad (13)$$

Proof: According to (1), C_1 is binomially distributed, that is, $\Pr\{C_1 = r\} = \binom{m}{r}P^r(1-P)^{m-r}$. Introducing the generating function $C(u) = \sum_{r=1}^{m} \Pr\{C_1 = r\}u^r$ for u complex, we easily get the formula $C(u) = (1 + P(u-1))^m$. Then, by Cauchy's celebrated formula [18]

$$\Pr\{C_1 \geq r\} = \frac{1}{2i\pi} \oint (1 + P(u-1))^m \frac{1}{u^r(u-1)} du , \quad (14)$$

where the integration is along a path around the unit disk for u complex. The problem is how to evaluate this integral for large m. In this case the best suited method seems to be the saddle point method [18, 13]. This method applies to integrals of the following form

$$I(m) = \int_{\mathcal{P}} \phi(x)e^{-mh(x)}dx , \quad (15)$$

where $\mathcal{P}$ is a closed curve, and $\phi(x)$ and $h(x)$ are analytical functions inside $\mathcal{P}$. We evaluate $I(m)$ for large m. The value of the integral does not depend on the shape of the curve $\mathcal{P}$. The idea is to run the path of integration through the saddle point, which is defined to be a place where the derivative of $h(x)$ is zero. To apply this idea to our integral (14) we represent it in the form of (15) and find the minimum of the exponent.

Define $u = 1 + h$, and then the exponent in our integral can be expanded as

$$\begin{aligned}
\log((1 + P(u-1))^m/u^r) &= m\log(1 + hP) - r\log(1 + h) \\
&= (Pm - r)h - 1/2(P^2m - r)h^2 + O(m + r)h^3) \\
&= -\tfrac{1}{2}(r - mP)^2(r - mP^2)^{-1} + (r - mP^2)\tfrac{(h-h_0)^2}{2} + O(m + r)h^3)
\end{aligned}$$

with $h_0 = (r - mP)/(r - mP^2)$. Let $r = mP + \sqrt{m}x$, with $x > 0$. Changing the scale of the variable under the integrand: $h = h_0 + it/\sqrt{m(P - P^2)}$, we obtain

$$\Pr\{C_1 \geq mP+\sqrt{m}x\} = \exp\left(\frac{x^2}{2(P - P^2 + x/\sqrt{m})}\right) \frac{1}{2\pi} \int_{-\log m}^{\log m} \frac{\exp(-t^2/2)}{\frac{x}{\sqrt{P-P^2}} + it} dt \,(1 + O(1/\sqrt{m})) .$$

Therefore, when $x \to \infty$ (e.g., as $\sqrt{\log m}$) we get

$$\int_{-\log m}^{\log m} \frac{\exp(-t^2/2)}{\frac{x}{\sqrt{P-P^2}} + it} dt \sim \frac{\sqrt{P - P^2}}{x} \int_{-\infty}^{\infty} \exp(-t^2/2)dt = \frac{\sqrt{2\pi(P - P^2)}}{x}$$

where the last integral can be computed from the error function [2]. This completes the proof of our lemma. ∎

To estimate the conditional probability $G(r,\sigma)$, we need some additional asymptotics for the binomial distribution, which are summarized in the lemma below.

Lemma 2 *Let* $r_m = mp + \sqrt{m\tau_m}$ *where* $\tau_m = O(\log m)$. *Then for all* $\rho > 1$ *there exists a sequence* $\gamma_m \to +\infty$ *such that the following hold*

$$\liminf_{m\to\infty} B(m - \gamma_m\sqrt{m}, p, r_m) \exp(\frac{\tau_m \rho}{2(p-p^2)}) \geq 1 \tag{16}$$

and

$$\limsup_{m\to\infty} B(m + \gamma_m\sqrt{m}, p, r_m) \exp(\frac{\tau_m}{\rho 2(p-p^2)}) \leq 1 \ . \tag{17}$$

Proof: These are computational consequences of Lemma 1. Details are deferred to the Appendix. ■

Now, we use Lemma 2 to estimate a typical behavior of the probability $G(r,\sigma)$ (recall that we treat $G(r,\sigma)$ as a random variable with respect to σ). For a string $\sigma \in \Sigma^m$, we define Ω_j as the number of occurrences of the symbol $j \in \Sigma$ in σ. If σ varies, then Ω_j can be treated as a random variable, too. Let also $C_1(j)$ be the number of matches that involve symbol of $j \in \Sigma$ between $\mathbf{b} = \sigma$ and $a_1 a_2 \cdots a_m$. Clearly, $C_1 = \sum_{j=1}^{V} C_1(j)$, and $C_1(j)$ is binomially distributed with parameters Ω_j and p_j. We have the following simple estimate of $G(r,\sigma)$.

Lemma 3 *For all* r_j *such that* $\sum_{j=1}^{V} r_j = r$, *the following holds*

$$\prod_{j=1}^{V} B(\Omega_j, p_j, r_j) \leq G(r,\sigma) \leq \sum_{j=1}^{V} B(\Omega_j, p_j, r_j) \ . \tag{18}$$

Proof: This follows immediately from the next two trivial bounds

$$\Pr\{C_1 > r\} \geq \Pr\{\forall j : \ C_1(j) > r_j\} \ ,$$

and

$$\Pr\{C_1 > r\} \leq \Pr\{\exists j : \ C_1(j) > r_j\} \ .$$

This completes the proof. ■

Our main result of this subsection is contained in the next theorem. It provides (a.s.) bounds for the probability $G(r,\sigma)$.

Theorem 3 *Let $r_m = mP + \sqrt{m\tau_m}$ with $\tau_m = O(\log m)$, and P and T as defined in Theorem 2. For all $\rho > 1$, the following estimates hold:*

(i)

$$\lim_{m\to\infty} \Pr\{\sigma :\ G(r_m,\sigma) \geq \exp\left(-\frac{\tau_m \rho}{2(P-T)}\right)\} = 1 \ , \tag{19}$$

(ii)

$$\lim_{m\to\infty} \Pr\{\sigma :\ G(r_m,\sigma) \leq \exp\left(-\frac{\tau_m}{\rho 2 S^2}\right)\} = 1 \ , \tag{20}$$

where $S = \sum_{j=1}^{V} p_j\sqrt{1-p_j}$.

Proof: We first deal with part (i). For convenience of presentation we drop the subscript m from r_m and τ_m. Let $r_j = mp_j^2 + \alpha_j\sqrt{m\tau}$ with $\sum_{j=1}^{V}\alpha_j = 1$. By (16) of Lemma 2, for any $\rho > 1$ there exists a sequence $\gamma_m \to \infty$ such that the following holds for m large enough

$$\prod_{j=1}^{V} B(mp_j - \gamma_m\sqrt{m}, p_j, r_j) \geq \prod_{j=1}^{V} \exp(-\frac{\alpha_j^2 \tau\rho}{2p_j^2(1-p_j)}) \ .$$

Taking $\alpha_j = p_j^2(1-p_j)/(\sum p_i^2(1-p_i))$, after some algebra involving an optimization of the right-hand side of the above, we obtain $(-\tau\rho/2(P-T))$ in the exponent of the right-hand side (RHS) of the above. But, by Lemma 3, $G(r,\sigma) \geq \prod_{j=1}^{V} B(mp_j - \gamma_m\sqrt{m}, p_j, r_j)$ where σ is such that for all j: $\Omega_j \geq mp_j - \gamma_m\sqrt{m}$. Thus,

$$\Pr\{\sigma :\ G(r,\sigma) \geq \exp(-\frac{\tau\rho}{2(P-T)})\} \geq \Pr\{\forall j :\ \Omega_j \geq mp_j - \gamma_m\sqrt{m}\} \ .$$

By the Central Limit Theorem, every random variable Ω_j tends to the Gaussian distribution with mean mp_j and with the standard deviation $\sqrt{m(p_j - p_j^2)}$. So, it is clear that

$$\lim_{m\to\infty} \Pr\{\forall j :\ \Omega_j \geq mp_j - \gamma_m\sqrt{m}\} = 1$$

as $\gamma_m \to \infty$.

The proof of part (ii) goes along the same lines. For completeness, we provide the details below. Let $r_j = mp_j^2 + \alpha_j\sqrt{m\tau}$ with $\sum_{j=1}^{V}\alpha_j = 1$. By Lemma 2, for every $\rho > 1$ we can find a sequence $\gamma_m \to \infty$ such that the following holds for m large enough:

$$\sum_{j=1}^{V} B(mp_j + \gamma_m\sqrt{m}, p_j, r_j) \leq V^{-1}\sum_{j=1}^{V} \exp(-\frac{\alpha_j^2\tau}{\rho 2p_j^2(1-p_j)})$$

Taking $\alpha_j = S^{-1}p_j\sqrt{1-p_j}$, we obtain $(-\tau/2S^2)$ in the exponent of the RHS of the above. But, by Lemma 3, $G(r,\sigma) \leq \sum_{j=1}^{V} B(mp_j + \gamma_m\sqrt{m}, p_j, r_j)$ where σ is such that for all j: $\Omega_j \leq mp_j + \gamma_m\sqrt{m}$. Thus,

$$\Pr\{\sigma :\ G(r,\sigma) \leq \exp\left(-\frac{\tau}{\rho 2S^2}\right)\} \geq \Pr\{\forall j :\ \Omega_j \leq mp_j + \gamma_m\sqrt{m}\} \ .$$

By the Central Limit Theorem, every Ω_j tends to the Gaussian random variable with mean mp_j and with standard deviation $\sqrt{m(p_j - p_j^2)}$. Hence,

$$\lim_{m\to\infty} \Pr\{\forall j:\ \Omega_j \geq mp_j + \gamma_m\sqrt{m}\} = 1$$

for $\gamma_m \to \infty$. ■

Remark 2. If we denote by $\mathcal{G}_m$ the set of $\sigma \in \Sigma^m$ such that $G(r,\sigma) \geq \exp\left(-\frac{\tau\rho}{2(P-T)}\right)$, then (19) asserts that $\Pr\{\mathcal{G}_m\} \to 1$. As above, if we denote by $\mathcal{G}'_m$ a set of $\sigma \in \Sigma^m$ such that $G(r,\sigma) \leq \exp(-\frac{\tau}{\rho 2S^2)})$, then (20) implies that $\Pr\{\mathcal{G}'_m\} \to 1$.

3.2 Upper Bounds

We establish here the two upper bounds claimed in Theorem 2. We prove the following theorem.

Theorem 4 *Let $\varepsilon > 0$ be an arbitrary non-negative real number. Then*

(i) If $\tau = 2(1+\varepsilon)(P - P^2)\log n$, then the following holds $\lim_{n,m\to\infty}\Pr\{M_{m,n} < r\} = 1$ with $r = mP + \sqrt{m\tau}$.

(ii) If $\tau = 2(1+\varepsilon)S^2\log n$, with $S = \sum_{i=1}^{V} p_i\sqrt{1-p_i}$, then $\lim_{n,m\to\infty}\Pr\{M_{m,n} < r\} = 1$ with $r = mP + \sqrt{m\tau}$.

Proof: Part (i) was in fact already proved in Theorem 1, and it really follows directly from $\Pr\{M_{m,n} < r\} \leq nB(m,P,r)$.

We concentrate on proving part (ii). We have

$$\begin{aligned}\Pr\{M_{m,n} > r\} &\leq \sum_{\sigma\in\mathcal{G}'_m} n\,p(\sigma)G(r,\sigma) + \Pr\{\sigma \notin \mathcal{G}'_m\} \\ &\leq \max_{\sigma\in\mathcal{G}'_m}\{nG(r,\sigma)\} + \Pr\{\sigma\notin\mathcal{G}'_m\}\ ,\end{aligned}$$

where the set $\mathcal{G}'_m$ is defined in Theorem 3 (ii) with $\rho = \sqrt{1+\varepsilon}$. Note that $\Pr\{\sigma \notin \mathcal{G}'_m\} \to 0$, by Theorem 3. Then by (20) $G(r,\sigma) \leq \exp(-\tau/(\rho 2S^2))$ for $\sigma \in \mathcal{G}'_m$, thus

$$\Pr\{M_{m,n} > r\} \leq n^{1-\sqrt{1+\varepsilon}} + \Pr\{\sigma \notin G_m\}\ .$$

But both terms of the above tend to zero, and this establishes the second upper bound in Theorem 2. ■

3.3 Lower Bounds

The lower bounds are much more intricate to prove. We start with a simple one that does *not* take into account overlapping. Then, we extend this bound to include the overlapping. This extension is of interest to us since it leads to the exact constant in the symmetric case.

In the first lower bound we ignore overlapping by considering $M^*_{m,n} \le M_{m,n}$ where $M^*_{m,n} = \max_{1\le i\le \lfloor n/m\rfloor}\{C_{1+im}\}$, that is, the maximum is taken over all nonoverlapping positions of $\mathbf{a}$. Note also that C_1 and C_{1+m} are *conditionally* independent (i.e., under $\mathbf{b}=\sigma$). Our first lower bound is contained in the theorem below.

Theorem 5 *Let* $0<\varepsilon<1$. *If* $\tau = 2(1-\varepsilon)(P-T)(1-\alpha)\log n$ *and* $r = mP+\sqrt{m\tau}$, *then* $\lim_{n,m\to\infty}\Pr\{M_{m,n}<r\}\to 0$.

Proof. Due to conditional independence we can write

$$\begin{aligned}\Pr\{M^*_{m,n}<r\} &= \sum_{\sigma\in\Sigma^m} p(\sigma)(1-G(r,\sigma))^{n/m}\\ &\le \max_{\sigma\in\mathcal{G}_m}\{(1-G(r,\sigma))^{n/m}\} + \Pr\{\sigma\notin\mathcal{G}_m\}\ ,\end{aligned}$$

where $\mathcal{G}_m$ is defined in remark 2, that is, for $\rho=(1-\varepsilon)^{-1/2}$ we have $G(r,\sigma)>\exp(-\tau\rho/(2(P-T)))$ for all $\sigma\in\mathcal{G}_m$. Note also that $\Pr\{\sigma\notin\mathcal{G}_m\}\to 0$. We concentrate now on the first term of the above. Using $\log(1-x)\le -x$, we obtain

$$(1-G(r,\sigma))^{n/m} \le e^{-(n/m)G(r,\sigma)}\ .$$

Now, it suffices to show that $(n/m)G(r,\sigma)\to\infty$ for $\sigma\in\mathcal{G}_m$. By Theorem 3, we have for $\sigma\in\mathcal{G}_m$

$$\frac{n}{m}G(r,\sigma) \ge \frac{n}{m}\exp\left(-\frac{\tau\rho}{2(P-T)}\right) \sim (n/m)^{1-\sqrt{1-\varepsilon}} \to \infty\ ,$$

where the convergence is a consequence of our restrictions $\log n = o(m)$. This completes the proof of the "easier" lower bound in Theorem 2. ∎

The second lower bound is more elaborate. It requires an estimate of $\Pr\{C_1>r\ \&\ C_i>r\}$ also for $i<m$. Let $F_{m,n}=\sum_{i=2}^m \Pr\{C_1>r\ \&\ C_i>r\}$. Note that $F_{m,n}\le F^*_{m,n} = \sum_{i=2}^m \Pr\{C_1+C_i>2r\}$. In the next subsection we prove the following theorem.

Theorem 6 *For* $\tau_m = O(\log m)\to\infty$ *we have*

$$F^*_{m,n}(mP+\sqrt{m\tau_m}) \sim \frac{m(P-3P^2+2T)^{5/2}}{2(T-P^2)\sqrt{\pi\tau_m^3}}\exp\left(-\frac{\tau_m}{P-3P^2+2T}\right)\ . \tag{21}$$

for $m\to\infty$. ∎

Assuming that Theorem 6 is available, we proceed as follows. Using the *second moment method* we have

$$\Pr\{M_{m,n} > r\} \geq \sum_{\sigma \in \Sigma^m} p(\sigma) S(r, \sigma) \ , \tag{22}$$

where following (9)

$$S(r,\sigma) = \frac{(nG(r,\sigma))^2}{nG(r,\sigma) + nF(r,\sigma) + (n^2 - n(2m+1))(G(r,\sigma))^2} \ .$$

In the above, the probability $F(r,\sigma)$ is defined in (12) (to recall, $F(r,\sigma) = \sum_{i=2}^{m} \Pr\{C_1 > r \ \& \ C_i > r \ | \mathbf{b} = \sigma\}$). We rewrite $S(r,\sigma)$ as

$$S(r,\sigma) = \left(\frac{1}{nG(r,\sigma)} + \frac{F(r,\sigma)}{n(G(r,\sigma))^2} + 1 - \frac{2m+1}{n} \right)^{-1} \ . \tag{23}$$

We also have the following identity

$$\sum_{\sigma \in \Sigma^m} p(\sigma) F(r,\sigma) = F_{m,n}(r) \ ,$$

where the probability $F_{m,n}(r) \leq F^*_{m,n}(r)$ is estimated in Theorem 6. The almost sure behavior of $F(r,\sigma)$ – as a random function of σ – is discussed in the next lemma.

Lemma 4 *If $\tau \to \infty$, then for any constant β we have*

$$\lim_{m\to\infty} \Pr\{\sigma : \ F(r,\sigma) \leq \beta m \exp\left(-\frac{\tau}{P - 3P^2 + 2T} \right) \} = 1 \ . \tag{24}$$

Proof: It is a simple consequence of Markov's inequality: Since $\sum_{\sigma \in \Sigma^m} p(\sigma) F(r,\sigma) = F_m(r)$, it is clear that $\Pr\{F(r,\sigma) > \gamma_m F_m(r)\} \leq 1/\gamma_m$. ■

Remark 3. We denote by $\mathcal{G}''_m$ the set of $\sigma \in \Sigma^m$ such that $F(r,\sigma) \leq \beta m \exp(-\frac{\tau}{P-3P^2+2T})$ holds. The lemma shows that $\lim_{m\to\infty} \Pr\{\sigma \in \mathcal{G}''_m\} = 1$. □

Now we are ready to prove the main result of this subsection that provides the most elaborate lower bound. Let δ_1 be defined as in Theorem 2, that is,

$$\delta_1 = \frac{(P-T)(P-3P^2+2T)}{3(T-P^2)} \ .$$

Let also $\delta_2 = \min\{(1-\alpha)\delta_1, 2(P-T)\}$.

Theorem 7 *Let $0 < \varepsilon < 1$. If $\tau = (1-\varepsilon)\delta_2 \log n$, then $\lim_{m,n\to\infty} \Pr\{M_{m,n} > r\} = 1$ for $r = mP + \sqrt{m\tau}$.*

Proof: Note that it suffices to prove the theorem for any positive and arbitrarily small ε. By (22) and (23) we need to show that

$$
\begin{aligned}
nG(r,\sigma) &\to \infty && (25)\\
\frac{F(r,\sigma)}{nG^2(r,\sigma)} &\to 0 \ . && (26)
\end{aligned}
$$

The first identity is easy to prove. For $\sigma \in \mathcal{G}_m$ by Theorem 3 with $\rho \le (1-\varepsilon)^{-1/2}$ we have

$$
nG(r,\sigma) \ge n^{1-1/2\sqrt{1-\varepsilon}\delta_2/(P-T)} \to \infty
$$

since $\delta_2/(2(P-T)) \le 1$.

Now, we deal with (26). Note that $\delta_1 > 0$ since $P^2 < T$. For $\sigma \in \mathcal{G}_m \cap \mathcal{G}''_m$ where $\mathcal{G}_m$ and $\mathcal{G}''_m$ are defined as in (respectively) Theorem 3 and Lemma 4 (with $\beta = 1$), we have

$$
\begin{aligned}
\frac{F(r,\sigma)}{nG^2(r,\sigma)} &\le \frac{m}{n}\exp\left(\frac{\tau\rho}{P-T} - \frac{\tau}{P-3P^2+2P}\right)\\
&\le \frac{1}{n^{1-\alpha-(1-\varepsilon)\delta_2/\delta_1}}\exp\left((\rho-1)\frac{\tau}{P-T}\right)
\end{aligned}
$$

We know that $1-\alpha-\delta_2/\delta_1(1-\varepsilon) \ge \varepsilon\delta_2/\delta_1$. Choosing $\rho - 1 = O(\varepsilon^2)$ in the above, we finally obtain

$$
\frac{F(r,\sigma)}{nG^2(r,\sigma)} \le \frac{1}{n^{1-\alpha-\delta_2/\delta_1(1-\varepsilon)+O(\varepsilon^2)}} \to 0 \ ,
$$

since ε can be arbitrary small.

Putting everything together, we have just proved that $S(r,\sigma) \to 1$ for all $\sigma \in \mathcal{G}_m \cap \mathcal{G}''_m$. But, by (22), Theorem 3 and Lemma 4

$$
\begin{aligned}
\Pr\{M_{m,n} > r\} &\ge \sum_{\sigma\in\Sigma^m} p(\sigma)S(r,\sigma)\\
&\ge \Pr\{\sigma \in \mathcal{G}_m \cap \mathcal{G}''_m\} \min_{\sigma\in\mathcal{G}_m\cap\mathcal{G}''_m}\{S(r,\sigma)\} \to 1 \ ,
\end{aligned}
$$

which completes the proof. ■

3.4 Proof of Theorem 6

In this subsection, we prove Theorem 6 concerning the asymptotic behavior of the probability $\Pr\{C_1 > r,\ C_i > r\}$ for $i < m$. In fact, we evaluate the generating function $H_{m,\ell}(u,v) = \sum_{r_1,r_2} \Pr\{C_1 = r_1 \ \&\ C_\ell = r_2\}u^{r_1}v^{r_2}$, and then compute the probability $\Pr\{C_1 = r_1 \ \&\ C_\ell = r_2\}$ through the Cauchy integral as it was done in Lemma 1.

Let $\mathbf{x}$ and $\mathbf{y}$ be column vectors of dimension V, that is, $\mathbf{x} = \{x_i\}_{i=1}^V$ and $\mathbf{y} = \{y_i\}_{i=1}^V$. We define the scalar product $\langle \mathbf{x}, \mathbf{y} \rangle$ by $x_1 y_1 + \cdots + x_V y_V$. Then, the next crucial theorem captures some important combinatorial properties of $\{C_1, C_\ell\}$ that allow to estimate the generating function $H_{m,2}(u,v)$ for $\ell = 2$, and finally $H_{m,\ell}(u,v)$ for any ℓ (cf. Theorem 9).

Now we are ready to establish a closed-form formula for the generating function $H_{m,2}(u,v)$. We construct a recurrence relationship between the distributions of $\{C_1(\mathbf{b}), C_2(\mathbf{b})\}$ and $\{C_2(\mathbf{b}'), C_3(\mathbf{b}')\}$ where $\mathbf{b}'$ is the suffix of $\mathbf{b}$ of length $m-1$. In the above we write $C_i(\mathbf{b})$ instead of C_i in order to show explicitly a dependency of C_i on the string $\mathbf{b}$. With this in mind, we can proceed to the following key theorem.

Theorem 8 *We have the identity $H_{m,2}(u,v) = \langle \mathbf{x}(u), \mathbf{A}^{m-1}(u,v)\mathbf{y}(v)\rangle$, where $\mathbf{A}(u,v)$ is a $V \times V$ square matrix whose generic element $a_{ij}(u,v)$ satisfies the following*

$$a_{ij}(u,v) = p_i(1 + p_i(v-1) + p_j(u-1))$$

when $i \neq j$, and

$$a_{ii}(u,v) = p_i(1 + p_i(uv-1))$$

for $i = j$. The row vectors $\mathbf{x}(u)$ and $\mathbf{y}(v)$ are defined as $\mathbf{x}(u) = \{1 + p_i(u-1)\}_{i=1}^V$ and $y(v) = \{p_i(1 + p_i(v-1))\}_{i=1}^V$.

Proof: Let us define a random variable Γ_i as the number of matches between string $\mathbf{a}$ and $\rho^i(\mathbf{b})$ *without* counting the eventual first matching at position i (recall that $\rho^i(\mathbf{b})$ is the shifted version of $\mathbf{b}$ by i positions on the cycle). For example, $\Gamma_1 = C_1$ if there is no matching at position 1, and $\Gamma_1 = C_1 - 1$ otherwise. Define next the generating function $P_{i,m}(u,v)$ as

$$P_{i,m}(u,v) = \sum_{r_1, r_2} \Pr\{\Gamma_1 = r_1 \;\&\; C_2 = r_2 \;\&\; \text{string } b \text{ starts with symbol } i\} u^{r_1} v^{r_2} ,$$

and let $\mathbf{P}_m(u,v)$ denote the row vector $\{P_{i,m}(u,v)\}_{i=1}^V$. Note that $\mathbf{P}_1(u,v) = \mathbf{y}(v)$ and $H_{m,2}(u,v) = \sum_{i=1}^{i=V}(1 + p_i(u-1))P_{i,m}(u,v)$, thus $H_{m,2}(u,v) = \langle \mathbf{x}(u), \mathbf{P}_m(u,v)\rangle$.

The most interesting fact that we prove next is the following relationship $\mathbf{P}_m(u,v) = \mathbf{A}(u,v)\mathbf{P}_{m-1}(u,v)$ when $m > 1$. A proof of this relies on building a recurrence relationship between $\{C_1(\mathbf{b}), C_2(\mathbf{b})\}$ and $\{C_2(\mathbf{b}'), C_3(\mathbf{b}')\}$, as explained above. Let i and j be the two first symbols of string $\mathbf{b}$ and let k be the second symbol of string $\mathbf{a}$. When $i \neq j$ we have $\Gamma_1(\mathbf{b}) = \Gamma_2(\mathbf{b}') + 1$ and $C_2(\mathbf{b}) = C_3(\mathbf{b}')$ if $k = j$, $\Gamma_1(\mathbf{b}) = \Gamma_2(\mathbf{b}')$ and $C_2(\mathbf{b}) = C_3(\mathbf{b}') + 1$ if $k = i$, and $\Gamma_1(\mathbf{b}) = \Gamma_2(\mathbf{b}')$ and $C_2(\mathbf{b}) = C_3(\mathbf{b}')$ otherwise. When $i = j$, we have

$\mathbf{a} = 01100101110101011110101110010111$
$\mathbf{b} = 01011001010111$
$\mathbf{b}' = 1011001010111$

$$\boxed{0}\ \boxed{1}\ 1\ \boxed{1}\ 0\ \boxed{0}\ 1\ 0\ 1\ \boxed{1}\ 1\ 0\ \boxed{1}\ 0\ 1\ 1\ 1\ 1\ 0\ 1\ 0\ 1\ 1\ 1\ 0\ 0\ 1\ 0\ 1\ 1\ 1$$
$$\boxed{0}\ \boxed{1}\ 0\ \boxed{1}\ 1\ \boxed{0}\ 0\ 1\ 0\ \boxed{1}\ 0\ 1\ \boxed{1}\ 1 \qquad C_1(\mathbf{b}fb) = 6,\ \Gamma_1(\mathbf{b}) = 5,\ C_2(\mathbf{b}') = 5,\ \Gamma_2(\mathbf{b}') = 4$$
$$0\ \boxed{1}\ 0\ 1\ 1\ 0\ \boxed{0}\ \boxed{1}\ 0\ \boxed{1}\ \boxed{0}\ \boxed{1}\ 1\ \boxed{1} \qquad C_2(\mathbf{b}fb) = 7,\ \Gamma_2(\mathbf{b}) = 7,\ C_3(\mathbf{b}') = 7,\ \Gamma_3(\mathbf{b}') = 6$$
$$0\ 1\ \boxed{1}\ 1\ 0\ 0\ 1\ \boxed{0}\ \boxed{1}\ 1\ \boxed{1}\ \boxed{0}\ \boxed{1}\ 0\ \boxed{1}\ 1\ 1\ 1\ 0\ 1\ 0\ 1\ 1\ 1\ 0\ 0\ 1\ 0\ 1\ 1\ 1$$

Figure 4: Illustration of the relationship between $\{C_1, C_2\}$ and $\{\Gamma_2, \Gamma_3\}$ (boxes show matches).

$\Gamma_1(\mathbf{b}) = \Gamma_2(\mathbf{b}')+1$ and $C_2(\mathbf{b}) = C_3(\mathbf{b}')+1$ if $k = i$, and $\Gamma_1(\mathbf{b}) = \Gamma_2(\mathbf{b}')$ and $C_2(\mathbf{b}) = C_3(\mathbf{b}')$ otherwise. This is illustrated in Figure 4. Since $\{C_2(\mathbf{b}'), C_3(\mathbf{b}')\}$ has the same distribution as $\{C_1(\mathbf{b}'), C_2(\mathbf{b}')\}$, we obtain the following identity

$$\frac{P_{i,m}(u,v)}{p_i} = (1 + p_i(uv-1))P_{i,m-1}(u,v) + \sum_{j\neq i}(1 + p_i(v-1) + p_j(u-1))P_{j,m-1}(u,v) \ ,$$

which proves our theorem. ∎

The next theorem extends Theorem 8 and give a formula for the generating function $H_{m,\ell}(u,v)$, which is of independent interest.

Theorem 9 *For all $q < m$ the following holds: $H_{m,1+q}(u,v) = (H_{h,2}(u,v))^{q-\ell}(H_{h+1,2}(u,v))^{\ell}$, where $h = \lfloor \frac{m}{q} \rfloor$ and $\ell = m - hq$.*

Proof. For $i \leq q$ define $\mathbf{b}_{(i)}$ as a subsequence of string $\mathbf{b}$ obtained by selecting the ith symbol of $\mathbf{b}$, then the $i+q$th, then $i+2q$th, and so forth. For $1 \leq i \leq \ell$, string $\mathbf{b}_{(i)}$ is of length $h+1$, whereas for $\ell+1 \leq i \leq q$ its length is h. We can do the same with string $\mathbf{a}$ and obtain subsequences $\mathbf{a}_{(1)}, \ldots, \mathbf{a}_{(q)}$.

Let $\langle \mathbf{a}, \rho^i(\mathbf{b}), \rho^j(\mathbf{b})\rangle$ be a new notation for the two dimensional row vector $[C_i, C_j]$, that is, it represents the number of matches between $\mathbf{a}$ and simultaneously $\rho^i(\mathbf{b})$ and $\rho^j(\mathbf{b})$. It is easy to see that $[C_1, C_{1+q}] = \langle \mathbf{a}_{(1)}, \rho^1(\mathbf{b}_{(1)}), \rho^2(\mathbf{b}_{(1)})\rangle + \cdots + \langle \mathbf{a}_{(q)}, \rho^1(\mathbf{b}_{(q)}), \rho^2(\mathbf{b}_{(q)})\rangle$. Note that $\langle \mathbf{a}_{(i)}, \rho^1(\mathbf{b}_{(i)}), \rho^2(\mathbf{b}_{(i)})\rangle$'s has the same distribution as $[C_1, C_2]$ when $\mathbf{b}_{(i)}$ is of length

$h+1$ for $i \leq \ell$, and the same distribution as $[C_1, C_2]$ when $\mathbf{b}_{(i)}$ is of length h for $\ell < i \leq q$. This finally proves the theorem. ■

Theorem 9 establishes a closed form formula for the generating function of the joint distribution $\Pr\{C_1 = r_1, C_\ell = r_2\}$. Therefore, in principle we can recover the probabilities $\Pr\{C_1 = r_1, C_\ell = r_2\}$ from $H_{m,\ell}(u,v)$ by Cauchy's formula, as we did in Lemma 1. The difficulty is that the generating function $H_{m,\ell}(u,v)$ is expressed in terms of matrix $\mathbf{A}(u,v)$, so we need some tools from linear algebra to apply the saddle point method. However, before we plunge into this, we should treat the symmetric case (i.e., $p_k = 1/V$) separately since, as the next lemma shows, it possesses a very special property.

Lemma 5 *In the symmetric case, for all $i \neq j$, the random variables C_i and C_j are pairwise independent.*

Proof: It suffices to prove that C_1 is independent of C_{1+q} for all $1 \leq q \leq m$. In the symmetric case the $a_{ij}(u,v)$'s are all identical and equal to $\frac{1}{V}(1+\frac{1}{V}(u-1+v-1))$ except when $i=j$ where $a_{ii}(u,v) = \frac{1}{V}(1+\frac{1}{V}(uv-1))$. Note that $\mathbf{y}(v)$ coincides with an eigenvector of the matrix $\mathbf{A}$ and $\mathbf{A}(u,v)\mathbf{y}(v) = (1+\frac{1}{V}(u-1))(1+\frac{1}{V}(v-1))\mathbf{y}(v)$, and therefore $H_{m,2}(u,v) = (1+\frac{1}{V}(u-1))^m(1+\frac{1}{V}(v-1))^m$. This last formula shows that C_1 and C_2 are independent. Applying Theorem 9 one concludes that also $H_{m,1+q}(u,v) = (1+\frac{1}{V}(u-1))^m(1+\frac{1}{V}(v-1))^m$. Therefore C_1 and C_{1+q} are also independent. ■

Corollary 2 *In the symmetric case we have $M_{m,n} - mP \sim \sqrt{2m(1/V - 1/V^2)\log n}$ when $m, n \to \infty$, given $\log n = o(m)$.*

Proof: We already know that the above is true when $\alpha < 1$ (cf. Theorem 2). Using the result stated in Lemma 5 about pairwise independence of the C_i's, we can rewrite inequality (9) as

$$\Pr\{M_{m,n} > r\} \geq \frac{(n\Pr\{C_1 > r\})^2}{n\Pr\{C_1 > r\} + (n^2 - n)(\Pr\{C_1 > r\})^2} \ . \tag{27}$$

It is clear that the RHS of the above tends to 1 if and only if $n\Pr\{C_1 > r\} \to \infty$. Referring to Lemma 1, it is clear that this occurs for $\tau = 2(1-\varepsilon)(P - P^2)\log n$ and any arbitrarily chosen $\varepsilon > 0$. ■

Hereafter, we concentrate on the asymmetric case. Let $\lambda(u,v)$ be the principal eigenvalue of the matrix $\mathbf{A}(u,v)$. Let $\xi(u,v)$ (resp. $\Pi(u,v)$) be the corresponding right (resp. left) eigenvector of $\mathbf{A}(u,v)$ such that $\langle \Pi(u,v), \xi(u,v) \rangle = 1$, that is, $A(u,v)\xi(u,v) = \lambda(u,v)\xi(u,v)$ and $\mathbf{A}^T(u,v)\Pi(u,v) = \lambda(u,v)\Pi(u,v)$ (cf. [25, 24]). We distinguish the following three cases:

1. When $u = v = 1$, $\lambda(1,1) = 1$, $\xi(1,1) = \mathbf{y}(1)$ and $\Pi(1,1) = \mathbf{x}(1)$, the other eigenvalues are null.

2. When $v = 1$, $\lambda(u,1) = 1 + P(u-1)$, $\xi(u,1) = \xi(1,1) = \mathbf{y}(1)$ and $\Pi(u,1) = 1/\lambda(u,1)\,\mathbf{x}(u)$, the other eigenvalues are null.

3. When $u = 1$, $\lambda(1,v) = \lambda(v,1) = 1 + P(v-1)$, $\xi(1,v) = \mathbf{y}(v)$ and $\Pi(1,v) = 1/\lambda(1,v)\,\mathbf{x}(1)$, the other eigenvalues are null.

It follows that the other eigenvalues are $O((u-1)(v-1))$, and therefore we immediately prove the following fact.

Corollary 3 $H_{m,2}(u,v)/\lambda^{m-1}(u,v) = \langle \mathbf{x}(u), \xi(u,v)\rangle \langle \Pi(u,v), \mathbf{y}(v)\rangle + O((u-1)^m(v-1)^m)$.

Proof: This is a classical property of the principal eigenvalue and follows from the Perron-Frobenius theorem (the interested reader is referred to [3, 24, 25] for details). ■

As a consequence of Corollary 3 we have the following important expansion of the generating function $H_{m,\ell}(u,v)$.

Corollary 4 *Let* $F_m(u,v) = \sum_{i=2}^{m} H_{m,i}(u,v)$. *We have*

$$\frac{F_m(u,v)}{(\lambda(u,v))^m} = \frac{a(u,v) - a^m(u,v)}{1 - a(u,v)} + O((u-1)(v-1)) \ , \tag{28}$$

with $a(u,v) = \langle \mathbf{x}(u), \xi(u,v)\rangle \langle \Pi(u,v), \mathbf{y}(v)\rangle / \lambda(u,v)$

Proof: From Corollary 3 and Theorem 9 we have the estimate

$$H_{m,1+q}(u,v) = (\lambda(u,v))^m (a(u,v))^q + O((u-1)^{h+1}(v-1)^{h+1})) \ .$$

Therefore

$$\frac{F_m(u,v)}{(\lambda(u,v))^m} = \sum_{q=1}^{m-1} (a(u,v))^q + \sum_{h=2}^{m} O((u-1)^{h+1}(v-1)^{h+1}) \ .$$

The proof is easily completed by summing the geometric series in the last expression. ■

The following two lemmas present more detailed Taylor's expansions of the principal eigenvalue of $\mathbf{A}(u,v)$ defined in Corollary 4. These expansions are next used in the saddle point method to obtain a sharp estimate of $F_{m,n}(r)$ around $r = mP + \sqrt{2mP(1-P)\log n}$.

Lemma 6 *The Taylor expansion of* $\lambda(u,v)$ *to the second order is* $1 + (u-1)P + (v-1)P + (u-1)(v-1)(2T - P^2)$, *with* $T = p_1^3 + \cdots + p_V^3$.

Proof. We know that $\lambda(u,v) = 1 + (u-1)P + (v-1)P + O((u-1)(v-1))$. We adopt the following notation. If $f(u,v)$ is a function of two variables u and v, then we denote by $f_u(u,v)$ (resp. $f_v(u,v)$) the partial derivative of $f(u,v)$ with respect to u (resp. to v). We have $\lambda = \langle \Pi, \mathbf{A}\xi \rangle$, where the variables (u,v) have been dropped in the last expression to simplify the presentation. Thus, $\lambda_u = \langle \Pi_u, \mathbf{A}\xi \rangle + \langle \Pi, \mathbf{A}_u\xi \rangle + \langle \Pi, \mathbf{A}\xi_u \rangle$. Since $\mathbf{A}\xi = \lambda\xi$, $\mathbf{A}^T\Pi = \lambda\Pi$, and since $\langle \Pi_u, \xi \rangle + \langle \Pi, \xi_u \rangle = 0$ (because we assume $\langle \Pi, \xi \rangle = 1$), we get $\lambda_u = \langle \Pi, \mathbf{A}_u\xi \rangle$. Substituting $u = 1$ in the last expression, we obtain the identity $\lambda(1,v) = (P + 2T(v-1) + \sum_{i=1}^{i=V} p_i^4 (v-1)^2)/(1 + P(v-1))$, and after some simple algebra the proof is completed. ■

Lemma 7 *The Taylor expansion of* $a(u,v)$ *to the second order is* $1-(u-1)(v-1)(T-P^2)$.

Proof. Easy computations give $a(u,1) = a(1,v) = 1$, therefore $a(u,v)-1$ is $O((u-1)(v-1))$. Differentiating twice $a(u,v)$ and setting $u = v = 1$ leads to a formula beginning with $\langle \Pi_{uv}, \xi \rangle + \langle \Pi, \xi_{uv} \rangle$ and ending with a linear combination of scalar products involving first partial derivatives of Π, ξ, $\mathbf{x}$ and $\mathbf{y}$. These first derivatives are already known since Π and ξ are completely determined when $u = 1$ or $v = 1$. For $\langle \Pi_{uv}, \xi \rangle + \langle \Pi, \xi_{uv} \rangle$, we differentiate twice both sides of $\langle \Pi, \xi \rangle = 1$ in order to get $\langle \Pi_{uv}, \xi \rangle + \langle \Pi, \xi_{uv} \rangle + \langle \Pi_u, \xi_v \rangle + \langle \Pi_v, \xi_u \rangle = 0$, which leads to a complete determination of $a_{uv}(1,1)$. ■

Lemmas 6 and 7 are crucial to apply Cauchy's formula in order to estimate $F_{m,n}(r)$ for $r > mP$. To do that, we can use the double Cauchy formula

$$F_{m,n}(r) = \frac{1}{(2i\pi)^2} \oint \oint F_m(u,v) \frac{\mathrm{d}u\mathrm{d}v}{u^r(u-1)v^r(v-1)} \; .$$

This kind of integration is rather unusual. Since $\Pr\{C_i > r, C_j > r\} \leq \Pr\{C_i + C_j > 2r\}$ we can estimate $F^*_{m,n}(r) = \sum_{i=2}^{m} \Pr\{C_1 + C_i > 2r\}$, which leads to a single integration

$$F^*_{m,n}(r) = \frac{1}{2i\pi} \oint F_m(u,u) \frac{1}{u^{2r}(u-1)} \mathrm{d}u \; .$$

Finally, we can prove the following asymptotics for the tail of $F_{m,n}(r)$, which establishes Theorem 6.

Proof of Theorem 6. We parallel the proof of Lemma 1. By Cauchy and (28) we have

$$F^*_{m,n}(r) \sim \frac{1}{2i\pi} \oint \frac{\lambda^m(u,u)(a(u,u) - a^m(u,u))}{u^{2r}(1-a(u,u))} \frac{\mathrm{d}u}{u-1} \; , \tag{29}$$

the integration path encircling the unit disk. Let $1+h=u$. Using Lemmas 18 and 19, we obtain the expansion

$$\begin{aligned}\log(\lambda^m(u,u)/u^{2r}) &= m\log(1+2hP+h^2(2T-P^2)+O(h^3))-2r\log(1+h)\\ &= -2(r-mP)h+(r+2mT-3mP^2)h^2+O((m+r)h^3)\\ &= -(r-mP)^2(r-m(3P^2-2T))^{-1}+\\ &+ (r-3mP^2+2mT)(h-h_0)^2+O((m+r)h^3)\ ,\end{aligned}$$

with $h_0=(r-mP)/(r-3mP^2+2mT)$. Let $r=mP+\sqrt{m}x$. Substituting $h=h_0+it/\sqrt{m}$, and using $1-a(1+h,1+h)=h^2(T-P^2)+O(h^3)$ we get the first estimate

$$\frac{1}{2i\pi}\oint\frac{\lambda^m(u,u)a(u,u)\mathrm{d}u}{u^{2r}(u-1)(1-a(u,u))}=$$

$$=\exp\left(-\frac{x^2}{P-3P^2+2T}\right)\frac{m}{2(T-P^2)\pi}\int\frac{\exp(-(P-3P^2+2T)t^2)}{(\frac{x}{P-3P^2+2T}+it)^3}\mathrm{d}t(1+O(1/\sqrt{m}))\ .$$

Since $x=O(\sqrt{\log n})$, we obtain

$$\frac{1}{2i\pi}\oint\frac{\lambda^m(u,u)a(u,u)\mathrm{d}u}{u^{2r}(u-1)(1-a(u,u))}\sim\frac{m(P-3P^2+2T)^{5/2}}{2(T-P^2)\sqrt{\pi}x^3}\exp\left(-\frac{x^2}{P-3P^2+2T}\right)\ . \tag{30}$$

It remains to evaluate the second term in (29), that is,

$$\frac{1}{2i\pi}\oint\frac{\lambda^m(u,u)a^m(u,u)\mathrm{d}u}{u^{2r}(u-1)(1-a(u,u))}\ . \tag{31}$$

Using the estimates from Lemmas 18 and 19 we find

$$\log(\lambda^m(u,u)a^m(u,u)/u^{2r})=-(r-mP)h+(r+mT-2mP^2)h^2+O((m+r)h^3)\ .$$

Hence (31) becomes

$$\frac{1}{2i\pi}\oint\frac{\lambda^m(u,u)a^m(u,u)\mathrm{d}u}{u^{2r}(u-1)(1-a(u,u))}\sim\frac{m(P-2P^2+T)^{5/2}}{2\sqrt{\pi}x^3}\exp\left(-\frac{x^2}{P-2P^2+T}\right)\ . \tag{32}$$

Since $P-2P^2+T<P-3P^2+2T$ in the asymmetric case, the exponent in (32) is larger than in (30), so (30) is the leading term in the asymptotic expansion of $F^*_{m,n}(r)$. This concludes the proof of Theorem 6. ■

APPENDIX

Proof of Lemma 2: We first prove the existence of a sequence γ_m such that (16) holds. Since for all p and r we have $B(0,p,r)=0$ and $\lim_{x\to\infty}B(x,p,r)=1$, we can find for the Bernoulli distribution a sequence γ_m such that

$$B(m-\gamma_m\sqrt{m},p,r_m)>\exp(-\frac{\tau_m\rho}{2(p-p^2)})\ , \tag{33}$$

and

$$B(m-\gamma_m\sqrt{m}-1,p,r_m) \le \exp(-\frac{\tau_m\rho}{2(p-p^2)}) < 1 \ . \tag{34}$$

Our aim is to prove that $\lim_{m\to\infty}\gamma_m = \infty$. Let us assume contrary that $\liminf_{m\to\infty}\gamma_m < \infty$, thus there exists γ and a subsequence of γ_m bounded from the above by the constant γ. For simplicity of the presentation, we assume that $\gamma_m < \gamma$ for all m. Then, due to monotonicity of $B(m,p,r)$ with respect to m, we have $B(m-\gamma_m\sqrt{m}-1,p,r_m) > B(m-\gamma\sqrt{m}-1,p,r_m)$. Furthermore, for $r_m = (m-\gamma\sqrt{m}-1)p + \sqrt{(m-\gamma\sqrt{m}-1)\tau_m^*}$ with

$$\tau_m^* = \frac{(\sqrt{\tau_m}+p\gamma)^2}{1-\gamma/\sqrt{m}} = \tau_m + O(\sqrt{\tau_m}) \ , \tag{35}$$

we obtain

$$B(m-\gamma\sqrt{m}-1,p,r_m) \sim \frac{1}{\sqrt{2\pi(p-p^2)\tau_m^*}}\exp\left(-\frac{\tau_m+O(\sqrt{\tau_m})}{2(p-p^2)}\right) \gg \exp\left(-\frac{\rho\tau_m}{2(p-p^2)}\right) \ , \tag{36}$$

where the last inequality holds for large m due to $\rho > 1$ and $\tau_m \to \infty$. Clearly, the above contradicts (34), hence, $\gamma_m \to \infty$, as needed.

The proof for the second inequality is similar and is only sketched. We set γ_m such that

$$B(m+\gamma_m\sqrt{m},p,r_m) < \exp(-\frac{\tau_m}{\rho 2(p-p^2)}) \tag{37}$$

and

$$B(m+\gamma_m\sqrt{m}+1,p,r_m) \ge \exp(-\frac{\tau_m}{\rho 2(p-p^2)}) \ . \tag{38}$$

If one assumes $\gamma_m < \gamma < \infty$, then

$$B(m+\gamma_m\sqrt{m}+1,p,r_m) < B(m+\gamma\sqrt{m}+1,p,r_m) \sim \frac{1}{\sqrt{2\pi(p-p^2)\tau_m^*}}\exp(-\frac{\tau_m^*}{2(p-p^2)}) \ , \tag{39}$$

with $\tau_m^* = \tau_m + O(\sqrt{\tau_m})$, which contradicts $B(m+\gamma_m\sqrt{m}+1,p,r_m) \ge \exp(-\tau_m/2\rho(p-p^2))$.

ACKNOWLEDGEMENT

The authors sincerely thank a referee who pointed out an error in an earlier version of the paper.

References

[1] K. Abrahamson, Generalized String Matching, *SIAM J. Comput.*, 16, 1039-1051, 1987.

[2] Abramowitz, M. and Stegun, I., *Handbook of Mathematical Functions,* Dover, New York (1964).

[3] Aldous, D., *Probability Approximations via the Poisson Clumping Heuristic*, Springer Verlag, New York 1989.

[4] Apostolico, A., Atallah, M., Larmore, L., and McFadin, S., Efficient Parallel Algorithms for String Editing and Related Problems, *SIAM J. Computing*, 19, 968-988, 1990.

[5] Arratia, R., Gordon, L., and Waterman, M., An Extreme Value Theory for Sequence Matching, *Annals of Statistics*, 14, 971-993, 1986.

[6] Arratia, R., and Waterman, M., The Erdös-Rényi Strong Law for pattern matching with a Given Proportion of Mismatches, *Annals of Probability*, 17, 1152-1169, 1989.

[7] Arratia, R., Gordon, L., and Waterman, M., The Erdös-Rényi Law in Distribution, for Coin Tossing and Sequence Matching, *Annals of Statistics*, 18, 539-570, 1990.

[8] Atallah, M., Jacquet, P., and Szpankowski, W., Pattern Matching with Mismatches: A Simple Randomized Algorithm and Its Analysis, *Proc. Combinatorial Pattern Matching*, Tuscon 1992.

[9] Chang, W.I., and Lawler, E.L., Approximate String Matching in Sublinear Expected Time, *Proc. 31st Ann. IEEE Symp. on Foundations of Comp. Sci.*, 116-124, 1990.

[10] Chung, K.L. and Erdös, P., On the Application of the Borel-Cantelli Lemma, *Trans. of the American Math. Soc.*, 72, 179-186, 1952.

[11] DeLisi, C., The Human Genome Project, *American Scientist*, 76, 488-493, 1988.

[12] Feller, W., *An Introduction to Probability Theory and its Applications,* Vol. II, John Wiley & Sons, New York (1971).

[13] Flajolet, P., Analysis of Algorithms, in *Trends in Theoretical Computer Science* (ed. E. Börger), Computer Science Press, 1988.

[14] Galambos, J., *The Asymptotic Theory of Extreme Order Statistics,* John Wiley & Sons, New York (1978).

[15] Z. Galil, Open Problems in Stringology, *Combinatorial Algorithms on Words* (Eds. A. Apostolico and Z. Galil), 1-8 (1984).

[16] L. Guibas and A. Odlyzko, Periods in Strings *Journal of Combinatorial Theory*, Series A, 30, 19-43 (1981).

[17] L. Guibas and A. W. Odlyzko, String Overlaps, Pattern Matching, and Nontransitive Games, *Journal of Combinatorial Theory,* Series A, 30, 183-208 (1981).

[18] Henrici, P., *Applied and Computational Complex Analysis*, vol. I., John Wiley& Sons, New York 1974.

[19] Jacquet, P. and Szpankowski, W., Autocorrelation on Words and Its Applications. Analysis of Suffix Trees by String-Ruler Approach, INRIA Technical report No. 1106, October 1989; submitted to a journal.

[20] Karlin, S. and Ost, F., Counts of Long Aligned Matches Among Random Letter Sequences, *Adv. Appl. Probab.*, 19, 293-351, 1987.

[21] Knuth, D.E., J. Morris and V. Pratt, Fast Pattern Matching in Strings, *SIAM J. Computing*, 6, 323-350, 1977.

[22] Lander, E., Langridge, R., and D. Saccocio, Mapping and Interpreting Biological Information, *Comm. of the ACM*, 34, 33-39, 1991.

[23] Louchard, G., and Szpankowski, W., String Matching: A Preliminary Probabilistic Results, Universite de Bruxelles, TR-217, 1991.

[24] Noble, B. and Daniel, J., *Applied Linear Algebra*, Prentice-Hall, New Jersey 1988

[25] Seneta, E., *Non-Negative Matrices and Markov Chains*, Springer-Verlag, New York 1981.

[26] Szpankowski, W., On the Height of Digital Trees and Related Problems, *Algorithmica*, 6, 256-277 (1992).

[27] M. Zuker, Computer Prediction of RNA Structure, *Methods in Enzymology*, 180, 262-288, 1989.