1993

A Probabilistic Analysis of a String Editing Problem and Its Variations

Guy Louchard

Wojciech Szpankowski
Purdue University, spa@cs.purdue.edu

Report Number:
93-078

A PROBABILISTIC ANALYSIS OF A STRING EDITING PROBLEM AND ITS VARIATION

Guy Louchard
Wojciech Szpankowski

CSD TR-93-078
December 1993
(Revised July 1994)

A PROBABILISTIC ANALYSIS OF A STRING EDITING PROBLEM AND ITS VARIATIONS*

July 8, 1994

Guy Louchard
Laboratoire d'Informatique Théorique
Université Libre de Bruxelles
B-1050 Brussels
Belgium

Wojciech Szpankowski[†]
Department of Computer Science
Purdue University
W. Lafayette, IN 47907
U.S.A.

Abstract

We consider a string editing problem in a probabilistic framework. This problem is of considerable interest to many facets of science, most notably molecular biology and computer science. A string editing transforms one string into another by performing a series of weighted edit operations of overall maximum (minimum) cost. The problem is equivalent to finding an optimal path in a weighted grid graph. In this paper, we provide several results regarding a typical behavior of such a path. In particular, we observe that the optimal path (i.e., edit distance) is almost surely (a.s.) equal to αn for large n where α is a constant and n is the sum of lengths of both strings. More importantly, we show that the edit distance is well concentrated around its average value. In the so called independent model in which all weights (in the associated grid graph) are statistically independent, we derive some bounds for the constant α. As a by-product of our results, we also present a precise estimate of the number of alignments between two strings. To prove these findings we use techniques of random walks, diffusion limiting processes, generating functions, and the method of bounded difference.

[*] A preliminary version of this paper was presented at *Combinatorial Pattern Matching* conference, Padova, 1993.

[†] Support for this research was provided in part by by NSF Grants CCR-9201078, NCR-9206315 and INT-8912631, in part by AFOSR Grant 90-0107, NATO Grant 0057/89, and Grant R01 LM05118 from the National Library of Medicine. This research was completed while the second author was visiting INRIA, Rocquencourt, France, and he wishes to thank INRIA (projects ALGO, MEVAL and REFLECS) for a generous support.

1. INTRODUCTION

String editing problem arises in many applications, notably in text editing, speech recognition, machine vision and, last but not least, molecular sequence comparison (cf. [36]). Algorithmic aspect of this problem has been studied rather extensively in the past (cf. [2], [30], [32], [33] and [36]). In fact, many important problems on words are special cases of string editing, including the *longest common subsequence* problem (cf. [1], [14]) and the problem of *approximate pattern matching* (cf. [12] and [34]).

In sequel we review the string editing problem, its importance, and its relationship to the longest path problem in a special grid graph.

Let **b** be a string consisting of ℓ symbols on some alphabet Σ of size V. There are three operations that can be performed on a string, namely *deletion* of a symbol, *insertion* of a symbol, and *substitution* of one symbol for another symbol in Σ. With each operation is associated a *weight* function. We denote by $W_I(b_i)$, $W_D(b_i)$ and $W_Q(a_i, b_j)$ the weight of insertion and deletion of the symbol $b_i \in \Sigma$, and substitution of a_i by $b_j \in \Sigma$, respectively. An *edit script* on **b** is any sequence ω of edit operations, and the total weight of ω is the sum of weights of the edit operations.

The *string editing problem* deals with two strings, say **b** of length ℓ (for ℓong) and **a** of length s (for $s$hort), and consists of finding an edit script ω_{max} (ω_{min}) of minimum (maximum) total weight that transforms **a** into **b**. The maximum (minimum) weight is called the *edit distance from* **a** *to* **b**, and its is also known as the Levenshtein distance. In molecular biology, the Levenshtein distance is used to measure similarity (homogeneity) of two molecular sequences, say DNA sequences (cf. [33]).

The string edit problem can be solved by the standard dynamic programming method. Let $C_{\max}(i,j)$ denote the maximum weight of transforming the prefix of **b** of size i into the prefix of **a** of size j. Then, (cf. [2], [30], [36]).

$$\begin{aligned} C_{\max}(i,j) = \max\{C_{\max}(i-1,j-1) + W_Q(a_i,b_j) \quad &, \quad C_{\max}(i-1,j) + W_D(a_i)\ , \\ &, \quad C_{\max}(i,j-1) + W_I(b_j)\} \end{aligned}$$

for all $1 \le i \le \ell$ and $1 \le j \le s$. We compute $C_{\max}(i,j)$ row by row to obtain finally the total cost $C_{\max} = C_{\max}(\ell, s)$ of the maximum edit script. A similar procedure works for the minimum edit distance.

The key observation for us is to note that interdependency among the partial optimal weights $C_{\max}(i,j)$ induce an $\ell \times s$ grid-like directed acyclic graph, called further a *grid graph*. In such a graph vertices are points in the grid and edges go only from (i,j) point

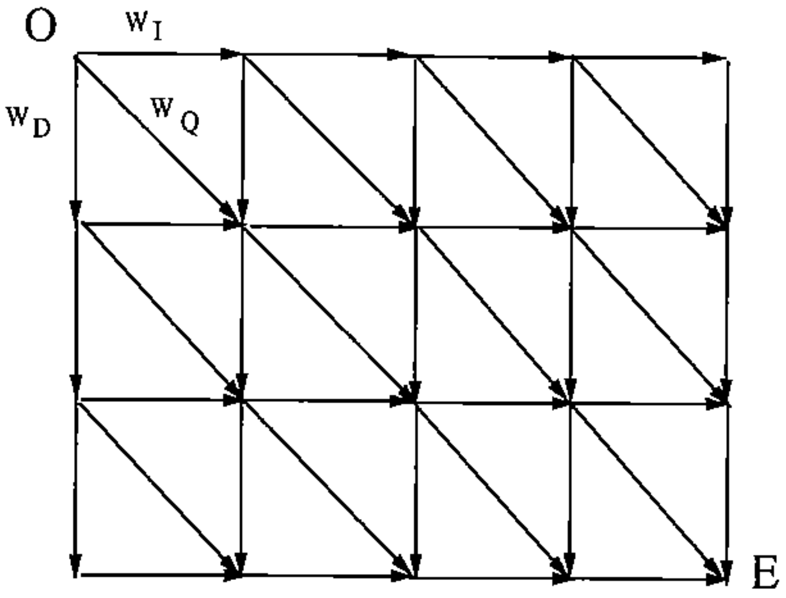

Figure 1: Example of a grid graph of size $\ell = 3$ and $s = 2$.

to neighboring points, namely $(i, j+1)$, $(i+1, j)$ and $(i+1, j+1)$. A horizontal edge from $(i, j-1)$ to (i, j) carries the weight $W_I(b_j)$; a vertical edge from $(i, j-1)$ to (i, j) has weight $W_D(a_i)$; and finally a diagonal edge from $(i-1, j-1)$ (i, j) is weighted according to $W_Q(a_i, b_j)$. Figure 1 shows an example of such an edit graph. The edit distance is the longest (shortest) path from the point $O = (0,0)$ to $E = (\ell, s)$.

In this paper, we analyze the string edit problem in a probabilistic framework. We adopt the Bernoulli model for a random string, that is, *all symbols of a string are generated independently with probability p_i for symbol $i \in \Sigma$.* A standard probabilistic model assumes that both strings are generated according to the Bernoulli scheme (cf. [3], [6], [7], [8], [14], [22], [35], [36]). We call it the **string model**. Such a framework, however, leads to statistical dependency of weights in the associated grid graph. To avoid this problem, most of the time we shall work within the framework of another probabilistic model which postulates that all weights in the associated grid graph are statistically independent. We call it **independent model**. This is closely related to a model in which *only* one string is random, say **b**, while the other one , say **a**, is deterministic. Indeed, in such a situation all weights in a "horizontal" strip in the associated grid graph are independent, while weights in a "vertical" strip are dependent (e.g., if **a** = 101, and **b** is random, then the "1"s in the string **a** match independently all "1"s in **b**, but clearly the first "1" and the third "1" in **a** have to match "1"s in **b** at the same places). We call such a model **semi-independent.**

Most of the results in this paper deal either with the independent model or the string model. We believe that better understanding of the independent model should be the first step to obtain valuable results for the semi-independent model. Certainly, results of the semi-independent model can be further used to deduce probabilistic behavior of the string model (cf. Theorem 2.2). In passing, we note that the semi-independent model might be useful in some applications (e.g., when comparing a given string to all strings in a data base).

In the independent model the distributions of weights $W_D(a_i)$, $W_I(b_j)$ and $W_Q(a_i, b_j)$ depend on the given string **a**. However, to avoid complicated notations we ignore this fact – whenever the independent model is discussed – and consider a grid graph with weights W_I, W_D and W_Q. In other words, we concentrate on finding the longest path in a grid graph with independent weights W_I, W_D and W_Q, not necessary equally distributed. By selecting properly these distributions, we can model several variations of the string editing problem. For example, in the standard setting the deletion and insertion weights are identical, and usually constant, while the substitution weight takes two values, one (high) when matching between a letter of **a** and a letter of **b** occurs, and another value (low) in the case of a mismatch (e.g., in the *Longest Common Substring* problem, one sets $W_I = W_D = 0$, and $W_Q = 1$ when a matching occurs, and $W_Q = -\infty$ in the other case).

Our results can be summarized as follows: Applying the *Subadditive Ergodic Theorem* we note that for the string model and the independent model $C_{\max} \sim \alpha n$ almost surely (a.s.), where $n = \ell + s$ (cf. Theorem 2.1 and Theorem 2.2). Our main contribution lies in establishing bounds for the constant α (cf. Theorem 2.7) for the independent model (cf. Theorem 2.2 for a possible extension to the string model). The upper bound is rather tight as verified by simulation experiments. More importantly, using the powerful and modern method of bounded differences (cf. [29]) we establish for all three models a sharp concentration of $C_{\max}$ around its mean value under a mild condition on the tail of the weight distributions (cf. Theorem 2.3). This proves the conjecture of Chang and Lampe [13] who observed empirically such a sharp concentration of $C_{\max}$ for a version of the string edit problem, namely the approximate string matching problem.

Our probabilistic results are proved in a unified manner by applying techniques of random walks (cf. [18], [20]), generating functions (cf. [19], [26], [27]), and bounded differences (cf. [29]). In fact, these techniques allow us to establish further results of a more general interest. In particular, we present an asymptotic estimate for the number of paths in the grid graph (cf. Theorem 2.4), which coincides with the number of sequence alignments (cf. [15], [16], [36]). Finally, for the independent model we establish the limiting distribution of the total weight (cf. Theorem 2.5) and the tail distribution of the total weight (cf. Theorem

2.6) of a randomly selected path (edit script) in the grid graph.

The string edit problem and its special cases (e.g., the longest common subsequence problem and the approximate pattern matching) were studied quite extensively in the past, and are subject of further vigorous research due to their vital application in molecular biology. There are many algorithmic solutions to the problem, and we only mention here Apostolico and Guerra [1], Apostolico *et al.* [2], Chang and Lampe [13], Myeres [30], Ukkonen [34], and Waterman [36]. On the other hand, a probabilistic analysis of the problem was initiated by Chvatal and Sankoff [14] who analyzed the longest common subsequence problem. After an initial success in obtaining some probabilistic results for this problem, and its extensions by a rather straightforward applications of the subadditive ergodic theorem, a deadlock was reached due to a strong interdependency between weights in the grid graph. To the best of our knowledge, there is no much literature on the probabilistic analysis of the string edit problem and its variations with a notable exception of a recent marvelous paper by Arratia and Waterman [7] (cf. [35]) who proved their own conjecture concerning phase transitions in a sequence matching.

There is, however, a substantial literature on probabilistic analysis of pattern matching. We mention here a series of papers by Arratia and Waterman (cf. [5], [6]) and with Gordon (cf. [3], [4]), as well as papers by Karlin and his co-authors (cf. [11], [21], [22]). Another approach for the probabilistic analysis of pattern matching with mismatches was recently reported by Atallah *et al.* in [8].

This paper is organized as follows. In the next section, we present our main results and discuss some of their consequences. Most of our proofs appear in Section 3.

2. MAIN RESULTS

We study a grid graph of size ℓ and s ($\ell \geq s$) as shown in Figure 1. All of our results, however, will be expressed in terms of $n = \ell + s$ and $d = \ell - s$. We assign to every edge in such a graph a real number representing its weight. A family of such directed acyclic weighted graphs will be denoted by $\vec{\mathcal{G}}(n,d)$ or shortly $\vec{\mathcal{G}}(n)$. We write $\vec{G}(n) \in \vec{\mathcal{G}}(n,d)$ for a member of such a family.

For the independent model we assume that *weights are independent* from edge to edge. Let $F_I(\cdot)$, $F_D(\cdot)$ and $F_Q(\cdot)$ denote distribution functions of W_I, W_D and W_Q respectively. We assume that the mean values m_I, m_D and m_Q, and the variances s_I^2, s_D^2 and s_Q^2, respectively, are finite. The distribution functions are not necessary identical.

The edit distance can be viewed as an optimization problem on the grid graph. Indeed, let $\mathcal{B}(n,d)$ or shortly $\mathcal{B}(n)$ be the set of all directed paths from the starting point O of the grid graph to the end point E. (It corresponds, as we know, to a script in the original string edit problem.) The cardinality of $\mathcal{B}(n)$, that is, the total number of paths between O and E, is denoted by $L(n,d)$. A particular path from O to E is denoted as $\mathcal{P}$, i.e., $\mathcal{P} \in \mathcal{B}(n,d)$. Note that the length $|\mathcal{P}|$ of a path $\mathcal{P}$ satisfies $\ell \le |\mathcal{P}| \le l + r = n$. Finally, let $N_I(\mathcal{P})$, $N_D(\mathcal{P})$ and $N_Q(\mathcal{P})$ denote the number of horizontal edges (say I-steps), vertical edges (say D-steps), and diagonal edges (say Q-steps) in a path $\mathcal{P}$.

With the above notation in mind, the problem at hand can be posed as follows:

$$C_{\max} = \max_{\mathcal{P}\in\mathcal{B}(n)}\{W_n(\mathcal{P})\} \quad , \quad C_{\min} = \min_{\mathcal{P}\in\mathcal{B}(n)}\{W_n(\mathcal{P})\} \tag{1}$$

where $W_n(\mathcal{P})$ denotes the total weight of the path $\mathcal{P}$ which becomes

$$W_n(\mathcal{P}) = \sum_{i=1}^{N_I(\mathcal{P})} W_I(i) + \sum_{i=1}^{N_D(\mathcal{P})} W_D(i) + \sum_{i=1}^{N_Q(\mathcal{P})} W_Q(i)\ . \tag{2}$$

We write W_n to denote the total weight of a *randomly* selected path, that is,

$$\Pr\{W_n < x\} = \frac{1}{L(n,d)} \sum_{\mathcal{P}\in\mathcal{B}} \Pr\{W_n(\mathcal{P}) < x\}\ . \tag{3}$$

Our results crucially depend on the order of magnitude of d with respect to n. We consider separately several cases. Below we define two of them that are analyzed in details in this paper:

CASE (A): $d = O(\sqrt{n})$, and let $x = d\sqrt{\sqrt{2}/n} = \zeta d/\sqrt{n}$ where $\zeta = 2^{1/4}$.
CASE (B): $d = \Theta(n)$, and let $x = d/n$.

Three other case, described below, are discussed in our extended technical report [28]:

CASE (C): $d = n - O(n^{1-\varepsilon})$, that is, for some constant x we have $d = n(1 - x/n^{\varepsilon})$.
CASE (D): $d = O(1)$ (we shall reduce this case to Case (A)).
CASE (E): $s = O(1)$ (we shall reduce this case to case (C)).

Now, we are in a position to present our results. To simplify further our presentation, we concentrate mainly on the longest path $C_{\max}$. We start with a simple general result concerning the typical behavior of $C_{\max}$. The more refined results containing a computable upper bound for $EC_{\max}$ (in the independent model) are given at the end of this section (cf. Theorem 2.7).

Theorem 2.1. *In the string model and the independent model, the following holds*

$$\lim_{n\to\infty} \frac{C_{\max}}{n} = \lim_{n\to\infty} \frac{EC_{\max}}{n} = \alpha \qquad (a.s.)\ , \tag{4}$$

provided ℓ/s has a limit as $n \to \infty$.

Proof. Let us consider the $\ell \times s$ grid with starting point O and ending point E (cf. Fig. 1). Call it *Grid*(O,E). We also choose an arbitrary point, say A, inside the grid so that we can consider two grids, namely *Grid*(O,A) and *Grid*(A,E). Actually, point A splits the edit distance problem into two subproblems with objective functions $C_{\max}(O,A)$ and $C_{\max}(A,E)$. Clearly, $C_{\max}(O,E) \geq C_{\max}(O,A) + C_{\max}(A,E)$. Thus, under our assumption regarding weights, the objective function $C_{\max}$ is superadditive, and direct application of *Superadditive Ergodic Theorem* (cf. [25]) proves our result. ■

Remark 1. Observe that we cannot directly apply subadditive ergodic theorem to semi-independent model since weights are not stationary in this case. However, using an inductive argument, one can obtain similar results as above for the semi-independent model. In particular, since for the semi-independent model $EC_{\max}$ is superadditive, we immediately prove that $EC_{\max} \sim \alpha\ell$. □

In the string and semi-independent models, weights depend on strings $\mathbf{a}$ and $\mathbf{b}$, hence the constant α is a function of $\mathbf{a}$ and $\mathbf{b}$. Furthermore, the string model can be reduced to the semi-independent model as follows. Let $\mathbf{a}$ be a given string (i.e., not random), and let $P(\mathbf{a})$ be the probability of $\mathbf{a}$ occurrence in our standard Bernoulli model (e.g., for the binary alphabet $\Sigma = \{a, b\}$ we have $P(\mathbf{a}) = p^{|a|}(1-p)^{|b|}$ where p is the probability of a occurrence, and $|a|$ ($|b|$) is the number of a's (b's) in the string $\mathbf{a}$). Let $\alpha_{\mathbf{a}}$ be the constant in the semi-independent model.

Theorem 2.2. *In the string model, the constant α can be estimated as follows*

$$\alpha = \sum_{\mathbf{a} \in \mathcal{H}} \alpha_{\mathbf{a}} P(\mathbf{a}) \tag{5}$$

where $\mathcal{H}$ is the set of all possible strings $\mathbf{a}$ of length s over the alphabet Σ.

Proof. Observe the following

$$\sum_{\mathbf{a} \in \mathcal{H}} \alpha_{\mathbf{a}} P(\mathbf{a}) = E\alpha_{\mathbf{a}} = E\left(\lim_{\ell\to\infty} \frac{EC_{\max}}{\ell} \cdot \frac{\ell}{n}\right) = \lim_{\ell\to\infty} \frac{E(EC_{\max})}{\ell} = \alpha \ ,$$

where the first equality is just definition of the expected value, the second follows from Remark 1, while the last is a simple consequence of the bounded convergence theorem and $\ell/n \to 1$. ■

Finally, for the string and independent models we can report the following finding concerning the concentration of the edit distance. It proves the conjecture of Chang and Lampe

[13]. The proof of this result uses a powerful *method of bounded differences* or Azuma's type inequality (cf. [29]).

Theorem 2.3. (i) *If all weights are bounded random variables, say* $\max\{W_I, W_D, W_Q\} \le 1$, *then for arbitrary* $\varepsilon > 0$ *and large* n

$$\Pr\{|C_{\max} - EC_{\max}| > \varepsilon EC_{\max}\} \le 2\exp(-\varepsilon^2 \alpha n) \ . \tag{6}$$

(ii) *If the weights are unbounded but such that for large* n, $W_{max} = \max\{W_I, W_D, W_Q\}$ *satisfies the following*

$$n\Pr\{W_{max} \ge n^{1/2-\delta}\} \le U(n) \tag{7}$$

for some $\delta > 0$ *and a function* $U(n) \to 0$ *as* $n \to \infty$, *then*

$$\Pr\{|C_{\max} - EC_{\max}| > \varepsilon EC_{\max}\} \le 2\exp(-\beta n^{\delta}) + U(n) \tag{8}$$

for any $\varepsilon > 0$ *and some* $\beta > 0$.

Proof: We consider only the string model. Part (i) is a direct consequence of the following inequality of Azuma's type (cf. [29]): *Let* X_i *be i.i.d. random variables such that for some function* $f(\cdot, \ldots, \cdot)$ *the following is true*

$$|f(X_1, \ldots, X_i, \ldots, X_n) - f(X_1, \ldots, X_i', \ldots, X_n)| \le c_i \ , \tag{9}$$

where $c_i < \infty$ *are constants, and* X_i' *has the same distribution as* X_i. *Then,*

$$\Pr\{|f(X_1, \ldots, X_n) - Ef(X_1, \ldots, X_n)| \ge t\} \le 2\exp(-2t^2/\sum_{i=1}^{n} c_i^2) \tag{10}$$

for some $t > 0$. The above technique is also called the method of bounded differences.

Now, for part (i) it suffices to set $X_i = b_i$ for $1 \le i \le \ell$, and $X_i = a_{i-\ell}$ for $\ell + 1 \le i \le n$, where a_i and b_i are the i symbols of the two strings **a** and **b**. Under our Bernoulli model, the X_i are i.i.d. and (9) holds, with $f(\cdot) = C_{\max}$. More precisely,

$$|C_{\max}(X_1, \ldots, X_i, \ldots, X_n) - C_{\max}(X_1, \ldots, X_i', \ldots, X_n)| \le \max_{1 \le i \le n} \{W_{max}(i)\} \ . \tag{11}$$

where $W_{max}(i)$ is the ith independent version of W_{max} defined in the theorem. Clearly, for part (i) we have $c_i = 1$, thus we can apply (10). Inequality (6) follows from the above and $t = \varepsilon EC_{\max} = O(n)$.

To prove part (ii), we start with (11). But, this time we have for some c

$$\begin{aligned}\Pr\{|C_{\max} - EC_{\max}| \ge t\} &= \Pr\{|C_{\max} - EC_{\max}| \ge t \ , \ \max_{1 \le i \le n} \{W_{max}(i)\} \le c\} \\ &+ \Pr\{|C_{\max} - EC_{\max}| \ge t \ , \ \max_{1 \le i \le n} \{W_{max}(i)\} > c\} \\ &\le 2\exp(-2t^2/nc^2) + n\Pr\{W_{max} > c\} \ .\end{aligned}$$

Set now $t = \varepsilon EC_{\max} = O(n)$ and $c = O(n^{1/2-\delta})$, then

$$\Pr\{|C_{\max} - EC_{\max}| \geq \varepsilon EC_{\max}\} \leq 2\exp(-\beta n^{\delta}) + n\Pr\{W_{max} > n^{1/2-\delta}\} ,$$

for some constant $\beta > 0$, and this implies (8) provided (7) holds. ■

Remark 2. Theorem 2.3 holds also for the semi-independent model if one replaces in the right-hand side of (6) $2\exp(-\varepsilon^2 n)$ by $2\exp(-\varepsilon^2 \ell)$ and set $EC_{\max} = O(\ell)$.

Hereafter, we investigate *only* the independent model. For this model, we have obtained several new results regarding the probabilistic behavior of (longest) path in a weighted grid graph.

The next result presents limiting distribution of the total weight defined in (2). Its proof is quite complicated, however, it applies only standard techniques. Therefore, at the referee request, we omit completely the proof of this theorem. It can be found in our extended technical report [28].

Theorem 2.4. *The limiting distribution of the total weight satisfies*

$$\frac{W_n - n\mu_W}{\sqrt{n}\sigma_W} \to \mathcal{N}(0,1) \tag{12}$$

where $\mathcal{N}(0,1)$ is the standard normal distribution, and

$$\begin{aligned} \mu_W &= m_I\mu_I + m_D\mu_D + m_Q\mu_Q \ , & (13) \\ \sigma_W^2 &= \mu_I s_I^2 + \mu_D s_D^2 + \mu_Q s_Q^2 + \tilde{\sigma}_Q^2(m_I + m_D - m_Q)^2 & (14) \end{aligned}$$

where $\mu_I = EN_I(\mathcal{P})$, $\mu_D = EN_D(\mathcal{P})$, $\mu_Q = EN_Q(\mathcal{P})$ and $\tilde{\sigma}_Q^2 = varN_Q(\mathcal{P})$. Explicit formulas for these quantities can be found in the next section. ■

Our next result enumerates the number of paths $L(n,d)$ in the grid graph. It is also of interest to some other problems since $L(n,d)$ represents the number of ways the string **a** can be transformed into **b**, and this problem was already tackled by others (cf. [15], [16], [24], [36]) in the case of equal length strings (i.e., $\ell = s$). The formulation of this result depends on a parameter $u = d/n$ that takes different values for case (A) and (B), that is:

CASE (A): Set $d = x\sqrt{n/\sqrt{2}}$. Then, $u = x/\sqrt{\sqrt{2}n} = x/(\zeta\sqrt{n})$.
CASE (B): Set $u = x$.

Theorem 2.5. *Let $L(u) = L(n,d)$ be the number of paths in a grid graph $\vec{G} \in \vec{\mathcal{G}}(n)$. Then,*

$$L(u) = \frac{C\psi_2(\beta_2(u))^n}{\beta_2(u)^{n(1+u)/2}\sqrt{2\pi n V(u)}}(1 + O(1/n)) \tag{15}$$

where

$$\beta_2(u) = \frac{1+3u^2+u\sqrt{8(u^2+1)}}{1-u^2} \ , \tag{16}$$

$$\psi_2(u) = \psi_2[\beta_2(u)] = \frac{2u\beta_2(u)}{\beta_2(u)-1-u(1+\beta_2(u))} \ , \tag{17}$$

and C is a constant that is found in Section 3 (cf. (79)). In the above, $V(u)$ is the variance obtained from the generating function $h(z)$ defined as $h(z) = \psi_2(z\beta_2(u))/\psi_2(\beta_2(u))$, that is, $V(u) = h''(1) - 0.25(1-u^2)$ where $h''(z)$ is the second derivative of $h(z)$. ■

For most of our computations, we only need the asymptotics of $L(u)$ in the following a less precise form

$$\log L(u) = n\rho(u) - 0.5\log n + O(1) \ , \tag{18}$$

where $\rho(u)$, for cases (A), (B) is respectively

$$\rho(u) = -\log(\sqrt{2}-1) \ , \tag{19}$$

$$\rho(u) = \log\psi_2(\beta_2(u)) - \frac{1+u}{2}\log\beta_2(u) \ . \tag{20}$$

The details of the above derivations can be found in Section 3.

Finally, in order to obtain an upper bound for the cost $C_{\max}$, we need an estimate on the tail distribution of the total weight W_n along a random path. Formula (2) suggests to apply Cramer's large deviation result (cf. Feller [18]) with some modifications (due to the fact that the total weight W_n as in (2) is a sum of *random* number of weights). To avoid unnecessary complications, we consider in details only two cases, namely:

(a) all weights are *identically* distributed with mean $m = m_I = m_D = m_Q$ and the *cumulant function* $\Psi(s) = \log Ee^{s(W-m)}$ for the common weight $W - m$;

(b) insertion weight and deletion weight are constant, say all equal to -1 (e.g., $W_I = W_D = -1$), and the substitution weight $W_Q - m_Q$ has the cumulant function $\Psi_Q(s) = \log Ee^{s(W_Q-m_Q)}$. Such an assignment of weights is often encountered in the string editing problem.

Theorem 2.6. (i) *In the case (a) of all identical weights, define s^* as the solution of*

$$a = \Psi'(s^*) \ , \tag{21}$$

for a given $a > 0$, and let

$$Z_0(a) = s^*\Psi'(s^*) - \Psi(s^*) \ , \tag{22}$$

$$E_1(a) = -(s^*m + \Psi(s^*)) \ , \tag{23}$$

$$E_2^2(a) = \frac{\tilde{\sigma}_Q^2 m^2 + 2\tilde{\sigma}_Q^2 ma + \tilde{\sigma}_Q^2(\Psi'(s^*))^2 + (1-\mu_Q)\Psi''(s^*)}{2(1-\mu_Q)\tilde{\sigma}_Q^2\Psi''(s^*)} \ , \tag{24}$$

where $\mu_Q = EN_Q$ and $\widetilde{\sigma}_Q^2 = var\ N_Q$. Then,

$$\begin{aligned}\Pr\{W_n &> (1-\mu_Q)(a+m)n\} \sim \\ &\frac{1}{2s^* E_2(a)\widetilde{\sigma}_Q\sqrt{\pi(1-\mu_Q)n\Psi''(s^*)}}\exp\left(-n(1-\mu_Q)Z_0(a)+n\frac{E_1^2(a)}{4E_2^2(a)}\right) \quad (25)\end{aligned}$$

(ii) *In case (b) of constant I-weights and D-weights, we define s^* as a solution of*

$$a = \Psi_Q'(s^*) \ , \qquad (26)$$

and let

$$\begin{aligned} Z_0(a) &= s^*\Psi_Q'(s^*) - \Psi_Q(s^*) \ , & (27)\\ E_1(a) &= s^*(m_Q+2) + 2s^*a^* - \Psi(s^*) \ , & (28)\\ E_2^2(a) &= \frac{\widetilde{\sigma}_Q^2(m_Q+2)(m_Q+2+2a^*+4s^*\Psi_Q''(s^*)) + \widetilde{\sigma}_Q^2 a^*(a^*+4s^*\Psi_Q''(s^*)) + \mu_Q\Psi_Q''(s^*)}{2\mu_Q\widetilde{\sigma}_Q^2\Psi_Q''(s^*)} & (29)\end{aligned}$$

Then,

$$\begin{aligned}\Pr\{W_n &> \mu_Q(a+\beta/\mu_Q)n\} \sim \\ &\frac{1}{2s^* E_2(a)\widetilde{\sigma}_Q\sqrt{\pi\mu_Q n\Psi_Q''(s^*)}}\exp\left(-n\mu_Q Z_0(a)+n\frac{E_1^2(a)}{4E_2^2(a)}\right) \ . \quad (30)\end{aligned}$$

where $\beta = 2\mu_Q + m\mu_Q - 1$. ■

Having the above estimates on the tail of the total cost of a path in the grid graph $\vec{G} \in \vec{\mathcal{G}}(n)$, we can provide a more precise information about the constant α in our Theorem 2.1, that is, we compute an upper bound $\overline{\alpha}$ and a lower bound $\underline{\alpha}$ of α for the independent model. We prove below the following result, which is one of our main finding.

Theorem 2.7 *Assume the independent model.*
(i) *Consider first the identical weights case (a) above. Let a^* be a solution of the following equation*

$$(1-\mu_Q)Z_0(a^*) = \rho + \frac{E_1^2(a^*)}{4E_2^2(a^*)} \ , \qquad (31)$$

where ρ is defined in (19)-(20), and Z_0, E_1 and E_2^2 are defined in (22)-(24). Then, the upper bound $\overline{\alpha}$ of α becomes

$$\overline{\alpha} = (1-\mu_Q)(a^*+m) + O(\log n/n) \ . \qquad (32)$$

In the case (b) of constant I and D weights, let a^* be a solution of the equation

$$\mu_Q Z_0(a^*) = \rho + \frac{E_1^2(a^*)}{4E_2^2(a^*)} , \tag{33}$$

where Z_0, E_1 and E_2^2 are as in (27-30). Then,

$$\overline{\alpha} = \mu_Q(a^* + \beta/\mu_Q) + O(\log n/n) , \tag{34}$$

where β is defined in Theorem 2.6(ii).

(ii) *The lower bound $\underline{\alpha}$ of α can be obtained from a particular solution to our optimization problem (1). In particular, we have*

$$\underline{\alpha} = \max\{\mu_W, \ell m_D + s m_I, \alpha_{gr}\} , \tag{35}$$

where α_{gr} is constructed from a greedy solution of the problem, that is,

$$n\alpha_{gr} = (\ell + s(1-p))m_{max} \tag{36}$$

where $p = \Pr\{W_Q > W_I \text{ and } W_Q > W_D\}$, and $m_{max} = E\max\{W_I, W_D, W_Q\}$.

Proof. We first prove part (i) provided Theorem 2.6 is *granted* (cf. Section 3 for the proof). Observe that by Boole's inequality we have for any real x

$$\Pr\{C_{\max} > x\} \le \sum_{\mathcal{P}\in\mathcal{B}} \Pr\{W_n(\mathcal{P}) > x\} = L(u)\Pr\{W_n > x\}$$

where the last equality follows from (3). We now consider only case (a). Let $\beta(a) = (1-\mu_Q)Z_0(a)$ and $\gamma(a) = E_1^2(a)/(4E_2^2(a))$. Then, by Theorem 2.5 and 2.6(i) we have

$$\Pr\{C_{\max} > (1-\mu_Q)(a+m)n\} \le O(1/n)\exp(n(\rho + \gamma(a) - \beta(a))) .$$

Setting in the above $a = a^*$ as defined in (31), we prove our result.

The lower bound can be established either by considering some particular paths $\mathcal{P}$ or applying a simple algorithm like a greedy one. The greedy algorithm selects in every step the most expensive edge, that is, the average cost per step is $m_{max} = E\max\{W_D, W_I, W_Q\}$. Let $p = \Pr\{W_Q > W_I\ ,\ W_Q > W_D\}$. Observe that if there are k D-steps, then necessarily, there are $s-k$ Q-steps. But, the number of Q-steps is binomially distributed with parameters p and s. Thus,

$$n\alpha_{gr} = m_{max}\sum_{k=1}^{s}(\ell+k)\binom{s}{k}p^{s-k}(1-p)^k = \ell + s(1-p)n_{max} ,$$

Table 1: Simulation results for exponentially distributed weights with means $m_I = m_D = m_Q = 1$ for case (B) with $d = 0.6n$.

ℓ	s	$\underline{\alpha}$	α_{sim}	$\overline{\alpha}$
200	50	1.588	1.909	2.45
400	100	1.588	1.808	2.45
600	150	1.588	1.899	2.45
800	200	1.588	1.926	2.45
1000	250	1.588	1.922	2.45

and this proves our result. ■

We compared our bounds for $C_{\max}$ with some simulation experiments. In the simulation we restricted our analysis to uniformly and exponentially distributed weights, and here we only report the latter results. They are shown in Table 1. It is plausible that the normalized limiting distribution for $C_{\max}$ is double exponential (i.e., $e^{-e^{-x}}$), however, the normalizing constants are quit hard to find.

The editing problem can be generalized, as it was recently done by Pevzner and Waterman [32] for the longest common subsequence problem. In terms of the grid graph, their generalization boils down to adding new edges in the grid graph that connect *no-neighboring* vertices. In such a situation our Theorem 2.1 may not hold. In fact, based on recent results of Newman [31] concerning the longest (unweighted) path in a general acyclic graph, we predict that a phase transition can occur, and $C_{\max}$ may switch from $\Theta(n)$ to $\Theta(\log n)$. This was already observed by Arratia and Waterman [7] for another string problem, namely, for the score in the pattern matching problem.

3. ANALYSIS THROUGH THE RANDOM WALK APPROACH

In this section, we only analyze the independent model. To recall, we consider an $\ell \times s$ grid graph with independent weights W_I, W_D and W_Q. We represent a path in the grid graph $\vec{G}$ as a random walk. First of all, it is convenient to append our $\ell \times s$ graph to a full $\ell \times \ell$ grid graph, with all steps possible, as shown in Figure 2. It should be noted that in our new representation, a Q-step is twice as long as I-step and D-step, and therefore the increments in such a random walk are *not* independent (e.g., after the first diagonal move, the second one comes with probability one).

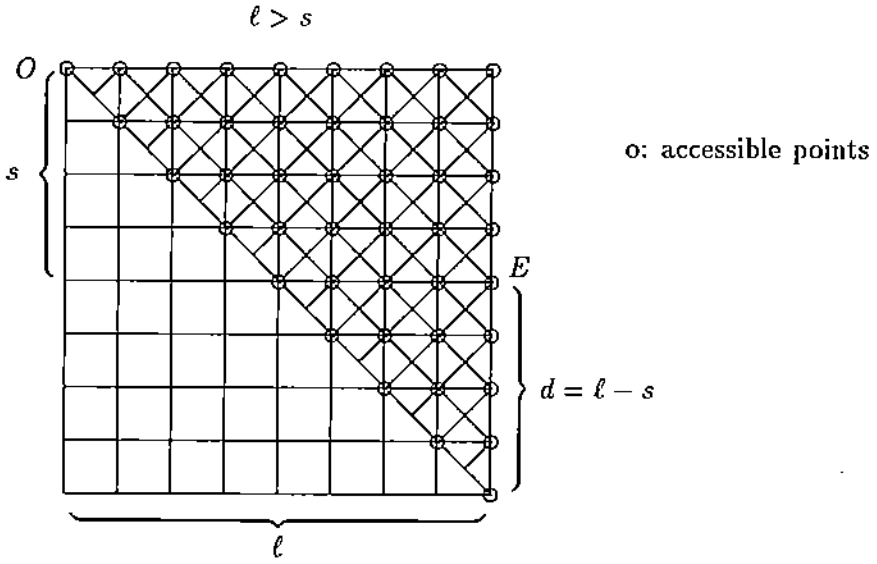

Figure 2: An extended $\ell \times \ell$ grid graph

We first analyze a path *without* weights in the grid graph shown in Figure 2. We call it an unweighted random walk (in short: R.W.) and denote as $Y(\cdot)$. To model a path $\mathcal{P}$ in our original problem, we must assure that the random walk $Y(\cdot)$ coincides with the script path $\mathcal{P}$, we require that the random walk $Y(\cdot)$ in Figure 2 ends at the point E of the grid graph after n steps where $n = 2\ell - (\ell - s) = \ell + s$. Thus, we impose the following constraint

$$Y(n) = d \tag{37}$$

where $d = \ell - s$.

We first consider an *unconstraint* random walk $\hat{Y}(\cdot)$ such that the condition (37) does *not* hold, and that the probabilities of I-step, D-step and Q-step $\tau = \sqrt{2} - 1$, τ and τ^2 respectively, as shown in Figure 3. These probabilities are chosen in such a way that all paths with the same length receive the same probability (e.g., a two-step path $I\&D$ has probability τ^2, the same as one-step path Q of length two).

3.1 Case (A): $d = O(\sqrt{n})$

Consider first the *unconstraint* random walk $\hat{Y}(\cdot)$ (cf. Fig. 3). We make the following scale changes $t = \frac{i}{n}$, $y = j\frac{\zeta}{\sqrt{n}}$ with $\zeta = \sqrt{\sqrt{2}}$ to establish the following theorem, where $\underset{n\to\infty}{\Longrightarrow}$ represents the weak convergence of random functions in the space of all right continuous functions having left limit and endowed with the Skorohod metric (see Billingsley [9] Ch.III).

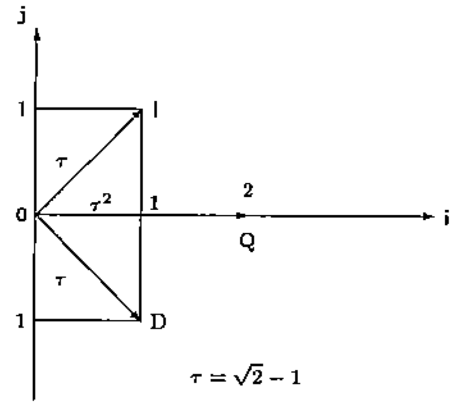

Figure 3: Probabilities of I-step, D-step and Q-step in the unconstraint random walk $\hat{Y}$

Theorem 3.1. *The unconstraint R.W. $\hat{Y}(\cdot)$ possesses the following limiting behavior*

$$\frac{\zeta \hat{Y}([nt])}{\sqrt{n}} \Rightarrow B(t), \qquad n \to \infty$$

where $B(\cdot)$ is a classical Brownian Motion (B.M.), and $\zeta = \sqrt{\sqrt{2}}$.

Proof. Let $p_i(j) = \Pr\{\hat{Y}(i) = j\}$. Then, $p_{i+1}(j) = \tau p_i(j-1) + \tau p_i(j+1) + \tau^2 p_{i-1}(j)$ for $i \geq 2$, and our result follows from standard arguments (cf. [28]). ∎.

Now, we take into account the constraint (37), that is, we set $Y(n) = d = O(\sqrt{n})$. Let

$$x = d\frac{\zeta}{\sqrt{n}} \tag{38}$$

with $\zeta = 2^{1/4}$. To handle this constraint, we recompute the probabilities of the I, D, and Q steps so that $EY(n) = d$ holds, and later we relax it so that our primary constraint (37) is true (cf. (48)). We define these new one-step probabilities as follows: $p_I = \Pr\{\text{first move is } I |\ Y(n) = d\}$, $p_D = \Pr\{\text{first move is } D |\ Y(n) = d\}$, and $p_Q = \Pr\{\text{first move is } Q |\ Y(n) = d\}$. Note that these probabilities depend on n, but we do not show explicitly this dependency.

Lemma 3.2a. *The new one-step probabilities become*

$$p_I = \tau\left(1 + \frac{\zeta}{\sqrt{n}}x\right) + O\left(\frac{1}{n}\right), \tag{39}$$

$$p_D = \tau\left(1 - \frac{\zeta}{\sqrt{n}}x\right) + O\left(\frac{1}{n}\right), \tag{40}$$

$$p_Q = \tau^2 + O\left(\frac{1}{n}\right). \tag{41}$$

Proof. We know from Theorem 3.1 that $f(t,x) = \frac{e^{-\frac{x^2}{2t}}}{\sqrt{2\pi t}}$. This, and the above definitions of p_I, p_D and p_Q lead to the following

$$p_I = \tau \frac{p_{n-1}(d-1)}{p_n(d)} = \frac{\tau}{p_n(d)} \left(p_n(d) - \frac{1}{n}\partial_t f - \frac{\zeta}{\sqrt{n}}\partial_x f + \frac{\zeta^2}{2n}\frac{\partial^2 f}{\partial x^2} + O\left(\frac{1}{n^{3/2}}\right)\right) .$$

But $\partial_x f = -\frac{x}{t} f$, hence p_I is now readily computed by setting $t = 1$ in the above. The two other probabilities are derived in a similar manner. ∎

Remark 3. In fact, using similar arguments to the ones in the proof of Theorem 3.1, we can prove much stronger result. Namely, the constraint random walk $Y(\cdot)$ characterized by the probabilities p_I, p_D, p_Q has the limiting density given by $f(y,v) = \exp(-\frac{(y-x)^2}{2v})/\sqrt{2\pi v}$, which is exactly the density of a B.M. with drift x and variance v.

To estimate the large deviation of the total weight W_n we need a precise evaluation of the random variables N_I, N_D and N_Q representing the number of I-steps, D-steps and Q-steps in a path $\mathcal{P}$. First of all, we compute the limiting distribution of the sum $N_I + N_D + N_Q$. Using the renewal theory (cf. Feller [18], p. 321, 341, and Iglehart [20] Theorem 4.1) we can easily prove that

$$N_I + N_D + N_Q \sim \mathcal{N}\left(\frac{n}{\bar{d}}, n\frac{\bar{\sigma}^2}{\bar{d}^3}\right) + O(1), \qquad n \to \infty \tag{42}$$

where $\mathcal{N}(m, \sigma^2)$ is a classical Gaussian variable with mean m and variance (VAR) σ^2. In the above, $\bar{d}$ is the average *move step*, that is, from Lemma 3.2 we have

$$\bar{d} = p_I + p_D + 2p_Q = 1 + p_Q \ , \tag{43}$$

so that $\bar{d} = 2(2-\sqrt{2}) + O(1/n)$, and

$$\bar{\sigma}^2 = p_Q(1-p_Q) \ , \tag{44}$$

hence $\bar{\sigma}^2 = \sqrt{2}(10 - 7\sqrt{2}) + O(1/n)$. Let

$$\alpha = \frac{1}{\bar{d}} = \frac{1}{1+p_Q} = \frac{2+\sqrt{2}}{4} + O\left(\frac{1}{n}\right) \tag{45}$$

and

$$\kappa = \frac{\bar{\sigma}^2}{\bar{d}^3} = \frac{\sqrt{2}}{16} + O\left(\frac{1}{n}\right) \tag{46}$$

Then, from (42), we obtain $N_I + N_D + N_Q \sim \mathcal{N}(n\alpha, n\kappa) + O(1)$.

From the expression (2) on the total weight W_n, it should be clear that we need the joint distribution of N_I, N_D and N_Q (cf. Louchard *et al.* [27]). For this, we must consider two constraints on $N.$:[1] one on the total number of steps, and the other related to $Y(n) = d$. More precisely, together with (38) we have the following constraint on the number of steps

$$\begin{aligned} N_I + N_D + 2N_Q &= n && (47) \\ N_I - N_D &= d = x\frac{\sqrt{n}}{\zeta} && (48) \end{aligned}$$

We first consider only the constraint (47). This will allow us to compute the asymptotic joint distribution of N_I, N_D, N_Q, as stated in the next theorem. The proof can be found in Appendix A.

Theorem 3.3a. *The number of I, D and Q steps, N_I, N_D, N_Q are asymptotically Gaussian, with mean $n\mu_I$, $n\mu_D$, $n\mu_Q$ respectively, where*

$$\begin{aligned} \mu_I &= \bar{p}_I(2\alpha - 1) = p_I/(1+p_Q) = \frac{\sqrt{2}}{4}\left(1 + \frac{\zeta}{\sqrt{n}}x + O(\frac{1}{n})\right) , && (49) \\ \mu_D &= \bar{p}_D(2\alpha - 1) = p_D/(1+p_Q) = \frac{\sqrt{2}}{4}\left(1 - \frac{\zeta}{\sqrt{n}}x + O(\frac{1}{n})\right) , && (50) \\ \mu_Q &= (1-\alpha) = p_Q/(1+p_Q) = \frac{2-\sqrt{2}}{4} + O(\frac{1}{n}) && (51) \end{aligned}$$

where $\bar{p}_I = p_I/(p_I+p_D)$, $\bar{p}_D = p_D/(p_I+p_D)$. Moreover, α is given by (45). The asymptotic covariance matrix is given by

$$n \cdot \begin{array}{c} I \\ D \\ Q \end{array} \begin{pmatrix} \sigma_I^2 & C_{ID} & -2\bar{p}_I\kappa \\ C_{ID} & \sigma_D^2 & -2\bar{p}_D\kappa \\ C_{IQ} & -2\bar{p}_D\kappa & \kappa \end{pmatrix} \qquad (52)$$

with

$$\begin{aligned} \sigma_I^2 &= (2\alpha-1)\bar{p}_I(1-\bar{p}_I) + 4\kappa\bar{p}_I^2 = \frac{3\sqrt{2}}{16} + \frac{2^{3/4}}{8}x + O(\frac{1}{n}) \\ \sigma_D^2 &= (2\alpha-1)\bar{p}_D(1-\bar{p}_D) + 4\kappa\bar{p}_D^2 = \frac{3\sqrt{2}}{16} - \frac{2^{3/4}}{8}x + O(\frac{1}{n}) \\ C_{ID} &= 2\kappa - (2\alpha-1)\bar{p}_I\bar{p}_D - 2\kappa(\bar{p}_I^2 + \bar{p}_D^2) = \frac{\sqrt{2}}{16} + O(\frac{1}{n}) \end{aligned}$$

where κ is given in (46). ■

To complete our study of the number of steps in the grid graph, we must take into account the constraint (48). Set $\eta. = (N. - n\mu.)/\sqrt{n}$. Observe that by Lemma 3.2a and Theorem 3.3a, $E(N_I - N_D) = \frac{\sqrt{n}}{\zeta}x + O(1)$ as it should be, so (48) and (47) imply respectively that $\eta_I = \eta_D$ and $\eta_I = -\eta_Q$.

[1] To simplify our notation, we often write $X.$ to denote any of X_I, X_D or X_Q.

To derive the *constrained* density of η_Q, we first write the joint asymptotic density $f(n_I, n_Q)$ of (η_Q, η_I), which by Theorem 3.3a becomes

$$f(n_I, n_Q) = \frac{\exp\left\{-\frac{1}{2(1-R^2)}\left(\frac{n_I^2}{\sigma_I^2} - 2R\frac{n_I n_Q}{\sigma_I \sigma_Q} + \frac{n_Q^2}{\sigma_Q^2}\right)\right\}}{2\pi\sigma_I\sigma_Q\sqrt{1-R^2}} \tag{53}$$

with $R = \frac{C_{IQ}}{\sigma_I \sigma_Q}$. Setting $\eta_I = -\eta_Q$, we finally obtain the asymptotic density of η_Q, as stated below.

Lemma 3.4a *Under constraint (48), we have*

$$\eta_Q \sim \mathcal{N}(0, \tilde{\sigma}_Q^2) + O(\frac{1}{\sqrt{n}}) \tag{54}$$

with

$$\tilde{\sigma}_Q^2 = (1 - R^2)\left(\frac{1}{\sigma_I^2} + 2\frac{C_{IQ}}{\sigma_I^2\sigma_Q^2} + \frac{1}{\sigma_Q^2}\right)^{-1} = \sqrt{2}/16 + O(1/n) \ , \tag{55}$$

where all the quantities in the above were defined before. ■

We delay the discussion of the number of paths $L(n, d)$ (cf. Theorem 2.5) until the next subsection since the recurrence on $L(n, d)$ is of the same kind as the one needed to study the behavior of W_n in the case (B). It will turn out that the asymptotics of $L(n, d)$ for (A) can be deduced from the asymptotics of $L(n, d)$ obtained in case (B).

Finally, we prove our last result concerning the large deviation of the total weight distribution (cf. Theorem 2.6). As discussed in Section 2, we only consider two cases, namely: (a) identically distributed weights, that is, $W_I =^d W_D =^d W_Q = W$ where $=^d$ means equal *in distribution*; and (b) constant D-weight and I-weight, i.e., $W_D = W_I = -1$.

Let us first establish notation needed to express a large deviation result. Define $S_n = \sum_{i=1}^n W(i)$ where $W(i)$ is an independent copy of W. Let $\Psi(z) = \log Ee^{z(W-m)}$ be the cumulant function of $W - m$ where $m = EW$, and let s be the unique solution, if exists, of the following equation

$$a = \Psi'(s)$$

for any $a > 0$. Finally, let $Z(a) = -(\Psi(s) - s\Psi'(s))$. Then (cf. Feller [18])

$$\Pr\{S_n \geq n(a + m)\} \sim \frac{1}{s\sqrt{2\pi n\Psi''(s)}} \exp(-nZ(a)) \ . \tag{56}$$

In our case, the total weight $W_n(N_Q)$ of a random path in a grid graph with exactly N_Q diagonal edges becomes $W_n(N_Q) = \sum_{i=1}^{N_I} W_I(i) + \sum_{i=1}^{N_D} W_D(i) + \sum_{i=1}^{N_Q} W_Q(i) = \sum_{i=1}^{n-N_Q} W(i)$ (cf. (47)). Note that N_Q is a random variable, hence the unconditional total weight W_n

can be computed from an estimate of the conditional total weight $W_n(N_Q)$ and the limiting distribution of N_Q (cf. Lemma 3.4a). But, $N_Q = n\mu_Q + \eta_Q\sqrt{n}$ and by Lemma 3.4a η_Q is asymptotically normal with mean 0 and variance $\tilde{\sigma}_Q^2$. We must now translate (56) into our new situation. Let $\tilde{n} = \gamma n$ where $\gamma = 1 - \mu_Q$. Define $\tilde{a}$ such that

$$\begin{aligned} \tilde{n}a + m\sqrt{n}\eta_Q &= (\tilde{n} - \sqrt{n}\eta_Q)\tilde{a}, \qquad \text{i.e.} \\ \tilde{a} &= a + \frac{(m+a)}{\gamma}\frac{\eta_Q}{\sqrt{n}} + \frac{m+a}{\gamma^2}\frac{\eta_Q^2}{n} + O\left(\frac{\eta_Q^3}{n^{3/2}}\right) \end{aligned}$$

Let also s^* and s be solutions of the following equations $a = \Psi'(s^*)$ and $\tilde{a} = \Psi'(s)$. Using Taylor's expansion of $\Psi(s)$ and $\Psi'(s)$ around s^*, we obtain

$$s = s^* + \frac{a^* + m}{\gamma\psi''(s^*)}\frac{\eta_Q}{\sqrt{n}} - \frac{1}{2}\frac{(a^*+m)[-2(\psi''(s^*))^2 + (a+m)\psi'''(s^*)]}{\gamma^2(\psi''(s^*))^3}\frac{\eta_Q^2}{n} + O(\frac{1}{n^{3/2}}) \,. \quad (57)$$

With the notation as above, we reduce the problem to the following one

$$\Pr\{W_n \geq \gamma(a+m)n\} = \int_{-\infty}^{\infty} \Pr\{\sum_{i=1}^{\tilde{n}-\sqrt{n}\eta}(W(i) - m) \geq (\tilde{n} - \sqrt{n}\eta)\tilde{a}|\eta_Q = \eta\}dF_{\eta_Q}(\eta)$$

where $F_{\eta_Q}(\eta) = \Phi(\eta)(1+O(1/\sqrt{n}))$ (cf. Lemma 3.4a) $\Phi(\cdot)$ stands for the normal distribution with mean zero and variance $\tilde{\sigma}_Q^2$. The probability under the above integral can be estimated as in (56). Using, in addition, the well known formula

$$\int_{-\infty}^{\infty} \exp(-p^2x^2 \pm qx)dx = \frac{\sqrt{\pi}}{p}\exp\left(\frac{q^2}{4p^2}\right) \,,$$

after tedious algebra, we obtain our result (25) presented in Theorem 2.6.

In a similar manner we deal with the second case (b). However, this time the starting equation is $W_n(N_Q) = \sum_{i=1}^{N_Q} W_Q(i) - (n - 2N_Q)$. The details are left to the reader.

3.2 Case (B): $d = O(n)$

The main purpose of this section is to derive the limiting distribution of the total weight for a given path $\mathcal{P}$ in a grid graph $\vec{G} \in \vec{\mathcal{G}}$, and the asymptotics for the number of paths $L(n,d)$. As in the previous subsection, we proceed in three steps: at first, we consider an unweighted unconstraint random walk, then we derive probabilities p_I, p_D and p_Q for the constraint unweighted random walk, and finally we deal with the total weight W_n.

Consider the unweighted random walk $Y(\cdot)$ in the grid graph as in Figure 2 such that $Y(n) = d = nx$ for some $x < 1$. Naturally, in this domain of d and n we can*not* use the normal approximation, which works only up to $O(\sqrt{n})$. We have to appeal to the large deviation arguments to obtain the probability distribution of the random walk $Y(\cdot)$. We proceed along the lines of arguments suggested by Louchard [26].

We consider the constraint random walk $Y(n) = nx$, however, it is convenient to generalize our constraint to the following one

$$Y(m) = mu \ . \tag{58}$$

One can imagine that the random walk $Y(\cdot)$ at step m has to be at position mu, where m and u are functions of n and x (e.g., we shall assume later that $mu = nx$).

As in the case (A), the analysis of the number of steps N_I, N_D and N_Q is crucial for the total weight. Note that, under our constraint (58), we have $N_I + N_D + 2N_Q = m$ and $N_I - N_D = mu$. The above can be translated to the following constraint: $N_I + N_Q = \frac{m}{2}(1+u)$. Bearing this in mind, we transformed the random walk $Y(\cdot)$ into another random walk $\tilde{Y}(\cdot)$ that is defined in Figure 4 below (i.e., its one-step moves are shown in Fig. 4). Our interest lies in estimating $\Pr\{Y(m) \in mdu\}$ or in terms of the new random walk $\tilde{Y}(\cdot)$

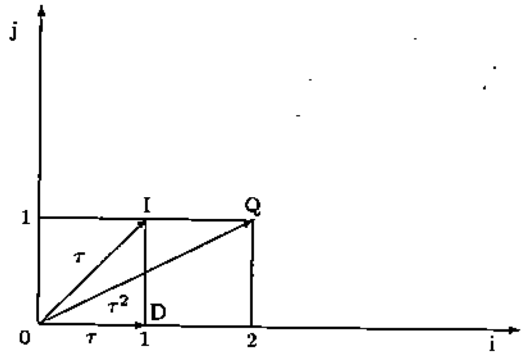

Figure 4: Definition of the new random walk $\tilde{Y}(\cdot)$.

we evaluate the following

$$\Pr\{Y(m) \in mdu\} \equiv \Pr\left\{\frac{\bar{Y}(m) - m/2}{m} \in \frac{du}{2}\right\} \tag{59}$$

To analyze $\bar{Y}(\cdot)$, we compute the probability $p_i(j) = \Pr\{\bar{Y}(i) = j\}$. It is easy to see that this probability satisfies the following recurrence

$$p_{i+1}(j) = \tau p_i(j) + \tau p_i(j-1) + \tau^2 p_{i-1}(j-1) \ , \qquad i \geq 1 \ . \tag{60}$$

We solve this recurrence by the mean of generating function approach. Let $g_i(z) = \sum_{j=0}^{\infty} z^j p_i(j)$. Clearly

$$g_0(z) \ = \ 1, \qquad g_1(z) = \tau(1+z) \tag{61}$$

$$g_{i+1}(z) \quad = \quad \tau g_i(z) + \tau z g_i(z) + \tau^2 z g_{i-1}(z), \qquad i \geq 1$$

Let now $\varphi(\theta, z) = \sum_{i=0}^{\infty} \theta^i g_i(z)$, and after some algebra one obtains

$$\varphi(\theta, z) = \frac{1}{1 - (1+z)\theta\tau - z\theta^2\tau^2} \tag{62}$$

The roots of the denominator of the above become

$$\theta_{1,2}(z) = -\frac{(1+z)\tau \pm \tau\sqrt{w_1(z)}}{2z\tau^2}$$

where $w_1(z) = 1 + 6z + z^2$. Then,

$$\varphi(\theta, z) = \left(\frac{\alpha_1(z)}{\theta - \theta_1(z)} + \frac{\alpha_2(z)}{\theta - \theta_2(z)} \right) . \tag{63}$$

where $\alpha_2(z) = -(\tau\sqrt{w_1(z)})^{-1}$, and $\alpha_1(z) = -\alpha_2(z)$.

To extract the generating function $g_i(z)$ from (63), we expand $\varphi(\theta, z)$ in the powers of θ to obtain

$$g_m(z) = -\frac{\alpha_1(z)}{\theta_1(z)} \left(\frac{1}{\theta_1(z)} \right)^m - \frac{\alpha_2(z)}{\theta_2(z)} \left(\frac{1}{\theta_2(z)} \right)^m \tag{64}$$

Since we are interested in large values of m, we deduce from (64) that the leading term of the asymptotics can be extracted from the following

$$g_m(z) \sim \frac{1}{\theta_2^m(z)} = \psi_1^m(z), \qquad m \to \infty \tag{65}$$

with $\psi_1(z) = 1/\theta_2(z)$. In the above, we omitted the function $\alpha_1(z)/\theta_1(z)$ since it only contributes a constant in the final asymptotics.

Our aim now is to assess asymptotically the probability $p_m(k) = \Pr\{\widetilde{Y}(m) = k\}$. Clearly, it can be estimated as $p_m(k) \sim [\psi_1^m(z)]_k$ where $[f(z)]_k$ is the coefficient of z^k in the power expansion of $f(z)$. Hence, we have to deal with evaluating the kth coefficient of of $\psi_1^m(z)$, where $k = m(1+u)/2$. To obtain such asymptotics we shall use the classical "shift of the mean" technique (cf. Feller [18] p.548 and Greene and Knuth [19] p.79). For the reader convenience, we discuss briefly this technique below. We follow the approach of Greene and Knuth [19].

Let $g(z)$ be the generating function of a random variable with mean equal to μ and the variance equal to σ^2. Then, $g^n(z)$ represents the generating function of the sum of n such independent random variables. We estimate the coefficient of $z^{\mu n + r}$ in $g^n(z)$ for such r that $\mu n + r$ is an integer. Call such a coefficient $A_{n,r}$. By the Cauchy formula Greene and Knuth [19] derive the following

$$A_{n,r} = \frac{1}{\sigma\sqrt{2\pi n}} \exp\left(\frac{-r^2}{2\sigma^2 n} \right) + O(n^{3\varepsilon - 1}) \tag{66}$$

where ε is arbitrary small positive number. The reader should notice that this asymptotics is valid *only* for $r = O(\sqrt{n})$.

In our case, we need the kth coefficient of $\psi_1^m(z)$, where $k = m(1+u)/2$. Therefore, we *cannot* directly apply (66) since we are not in the range $O(\sqrt{m})$. A solution to this dilemma is proposed in [19] by a simple and elegant application of the "shift of the mean" technique, which we discuss below.

Let us return to Greene and Knuth [19], and assume that one needs the kth coefficient of $g^n(z)$. The shift of the mean technique computes the kth coefficient as follows

$$[g^n(z)]_k = \frac{g(\beta)^n}{\beta^k}\left[\left(\frac{g(\beta z)}{g(\beta)}\right)^n\right]_k , \tag{67}$$

where the parameter β allows to shift the mean of the distribution to a value close to k/n, and hence allows to apply the asymptotics (66). The choice of β is specified by the following equation

$$\frac{\beta g'(\beta)}{g(\beta)} = \frac{k}{n} . \tag{68}$$

Now, we are ready to derive our asymptotics. Since we seek the $k = m(1+u)/2$ coefficient of $\psi_1^m(z)$, we first apply (68) to shift the mean. Define $\beta_1(u)$ as

$$\frac{\beta_1\psi_1'(\beta_1)}{\psi_1(\beta_1)} = \frac{1+u}{2} \tag{69}$$

Finally, applying (66), we obtain our main result.

Theorem 3.1b *We have proved*

$$\begin{aligned}\Psi_1(m,u) &= \Pr\{Y(m) \in mdu\} = \Pr\{\bar{Y}(m) - \frac{m}{2} \in \frac{mdu}{2}\} \\ &\sim \frac{(\psi_1(\beta_1))^m}{\beta_1^{m(1+u)/2}\sqrt{2\pi m V(u)}}\,\frac{mdu}{2}\,(1+O(1/n))\end{aligned}$$

for all m, where

$$\frac{(\beta_1(u)-1)\psi_1(u)}{2\tau\beta_1(u)+(1+\beta_1(u))\psi_1(u)} = u \tag{70}$$

$$\beta_1(u) = \frac{1+3u^2+u\sqrt{8(u^2+1)}}{1-u^2} \tag{71}$$

$$\psi_1(u) = \psi_1[\beta_1(u)] = \frac{2u\tau\beta_1(u)}{\beta_1(u)-1-u(1+\beta_1(u))} \tag{72}$$

for all $u < 1$. ■

Theorem 3.1b allows to analyze the constraint random walk $Y(n) = d$. In particular, as for case (A), we can compute the probabilities p_I, p_D, and p_Q of one-step moves. Setting in Theorem 3.1b, $m = nt, u = \frac{x}{t}$ so that $mu = nx$, we obtain

$$\begin{aligned}\Psi_2(x,t)dx &= \Pr\{Y(nt) \in ndx\} \\ &\sim \exp\left\{nt[\log \psi_1(\frac{x}{t}) - \frac{1}{2}\log\beta_1(\frac{x}{t})] - \frac{nx}{2}\log\beta_1(\frac{x}{t})\right\} \frac{\sqrt{n}\,\lambda[\beta_1(\frac{x}{t})]}{2\sqrt{2\pi t V(\frac{x}{t})}} \cdot dx\end{aligned}$$

This implies, for example, that

$$p_I \sim \frac{1}{3}\left[\frac{\Psi_2(x-\frac{1}{n}, t-\frac{1}{n})}{\Psi_2(x,t)}\right]_{t=1}$$

and in a similar fashion for D and Q. After some algebra, we finally derived the following lemma.

Lemma 3.2b. *The probabilities p_I, p_D and p_Q become*

$$\begin{aligned}p_I &= \frac{\tau\beta_1(x)}{\psi_1(x)} + O(1/n) \\ p_D &= \frac{\tau}{\psi_1(x)} + O(1/n) \\ p_Q &= \frac{\tau^2\beta_1(x)}{\psi_1^2(x)} + O(1/n)\end{aligned}$$

for all $x < 1$. ■

Concerning the limiting joint distribution of the number of I-steps, D-steps and Q-steps, we proceed as before. We use the same notation as in Theorem 3.3 with appropriate values for probabilities p_I, p_D and p_Q from Lemma 3.2b. This leads to the following results.

Theorem 3.3b. *The number of I, D and Q steps, N_I, N_D, N_Q respectively, are asymptotically Gaussian, with mean $n\mu_I$, $n\mu_D$, $n\mu_Q$ respectively, where these quantities are computed according to (49)-(51), (52) with new probabilities p_I, p_D and p_Q, as in Lemma 3.2b.* ■

Lemma 3.4b *We have $\eta_Q = \mathcal{N}(0, \tilde{\sigma}_Q^2)$ with $\tilde{\sigma}_Q^2$ given by (55) with probabilities p_I, p_D and p_Q as in Lemma 3.2b.* ■

Finally, we prove Theorem 2.5 that enumerates the total number of path $L(u)$. As discussed in Section 2, this estimate is necessary to evaluate our upper bound $\overline{\alpha}$ in Theorem 2.7. We start the analysis with setting up a recurrence for $L(u)$. Let $f_i(j)$ be the total number of paths from O to j in i steps of the associated random walk in our grid graph

$\vec{G}$. Then, $L(u) = f_n(d)$. Hereafter, we set $d = un$. Clearly, $f_i(j)$ satisfies the following recurrence

$$f_{i+1}(j) = f_i(j) + f_i(j-1) + f_{i-1}(j-1) \tag{73}$$

with $f_1(1) = 1$. This recurrence was already studied by Laquer [24] for $d = O(n)$. Observe that the above recurrence is similar to the one consider before, and we can use the same technique to attack it. Set $g_i(z) = \sum_{j=0}^{\infty} z^j f_i(j)$ and let $\varphi(\theta, z) = \sum_{i=0}^{\infty} \theta^i g_i(z)$. After the some algebra we obtain

$$\varphi(\theta, z) = \frac{\alpha_1(z)}{\theta - \theta_1(z)} + \frac{\alpha_2(z)}{\theta - \theta_2(z)} \tag{74}$$

with $w_2(z) = 1 + z^2 + 6z$ and

$$\theta_{1,2}(z) \quad = \quad \frac{1 + z \pm \sqrt{w_2(z)}}{-2z}, \tag{75}$$

where $\alpha_2(z) = -1/\sqrt{w_2(z)}$ and $\alpha_1(z) = -\alpha_2(z)$. As n becomes larger, the dominant contribution comes from (74), and asymptotically we have

$$g_n(z) \sim \lambda(z) \left[\frac{1}{\theta_2(z)}\right]^n = \lambda(z)\psi_2^n(z)$$

with $\lambda(z) = -\alpha_2(z)/\theta_2(z)$.

To extract the coefficient of $g_n(z)$ we shall apply the "shift of the mean" method, as described before. We first consider only the coefficient at $g_n(z)/\lambda(z) = \theta_2^{-n}(z)$. Call it $l(u)$. Applying equation (67), as in (69), we estimate the new mean value with $\psi_1(z)$ replaced by $\psi_2(z)$ and the new $\beta_2(u)$ becomes

$$\beta_2(u) = \frac{1 + 3u^2 + u\sqrt{8(u^2+1)}}{1 - u^2} \tag{76}$$

and then

$$\psi_2(u) = \psi_2[\beta_2(u)] = \frac{2u\beta_2(u)}{\beta_2(u) - 1 - u(1 + \beta_2(u))} \ .$$

Let $V(u)$ be the variance related to the generating function $\frac{\psi_2(\beta_2 z)}{\psi_2(\beta_2)}$ as defined in Theorem 2.5. With this in mind, it is easy to see that

$$l(u) = \frac{\psi_2(\beta_2(u))^n}{\beta_2(u)^{n(1+u)/2}\sqrt{2\pi n V(u)}}(1 + O(1/n)) = \frac{\exp(n\rho(u))}{\sqrt{2\pi n V(u)}}(1 + O(1/n)) \tag{77}$$

where $\rho(u)$ is a function of u and it is given by (20).

Now, we compute the coefficient at $g_n(z) = \lambda(z)\theta_2^{-n}(z)$, that is, we include the correction coming from $\lambda(z)$. Note that $\lambda_1(z) = \lambda(z)/\lambda(1)$ can be viewed as the generating function of a random variable. Let its probability distribution be denoted by $p_\lambda(i)$. Since the product of two generating functions translates into the convolution of the appropriate coefficients, we have $L(u) = \sum_{i=0}^{\infty} p_\lambda(i) l(u - i/n)$. By (77) we finally obtain

$$\begin{aligned} L(u)\sqrt{2\pi n V(u)} &= \lambda(1)\sum_{i=0}^{\infty} p_\lambda(i) \exp\left(n(\rho(u) - \rho'(u)i/n + O(n^{-2}))\right)(1 + O(n^{-1})) \\ &= \lambda(e^{-\rho'(u)})\exp(n\rho(u))(1 + O(n^{-1})) \ , \end{aligned} \tag{78}$$

where $\rho'(u)$ is the derivative of $\rho(u)$. From the above, we conclude that the constant C in Theorem 2.5 becomes

$$C = \lambda(e^{-\rho'(u)}) \ . \tag{79}$$

This completes the proof of Theorem 2.5 for case (A) and (B) (in case (A) $\rho(u)$ is given by (19)).

APPENDIX A: Proof of Theorem 3.3a

From (42) and (47) after setting $N_T = N_I + N_D$, we see that

$$\frac{n}{2} + \frac{N_T}{2} = n - N_Q \sim \mathcal{N}(n\alpha, n\kappa) + O(1) \tag{80}$$

$$\begin{aligned} \mu_T &= E(N_T)/n = n(2\alpha - 1) + O(1) \\ \sigma_T^2 &= \mathrm{VAR}(N_T)/n \sim 4\kappa \end{aligned} \tag{81}$$

But *given* N_T, the number of I-steps N_I is a binomial random variable with parameter $\bar{p}_I$, and mean $N_T\bar{p}_I$ and the variance $N_T\bar{p}_I\bar{q}_I$ (where $\bar{q}_I = 1 - \bar{p}_I$). By (81) we have $E(N_I) = n\mu_T\bar{p}_I = n\bar{p}_I(2\alpha - 1)$. We also obtain $E(N_I^2|N_T) = N_T\bar{p}_I\bar{q}_I + N_T^2\bar{p}_I^2$ and $E(N_I^2) = n\mu_T\bar{p}_I\bar{q}_I + \bar{p}_I^2(n\sigma_T^2 + n^2\mu_T^2)$. This finally leads to

$$\sigma_I^2 = n\left(\mu_T\bar{p}_I\bar{q}_I + \bar{p}_I^2\sigma_T^2\right) \sim n((2\alpha - 1)\bar{p}_I\bar{q}_I + \bar{p}_I^2 4\kappa)$$

The number of D-steps is analyzed in a similar fashion. To compute the covariance C_{ID} between N_I and N_D, note that

$$n\sigma_T^2 = \mathrm{VAR}(N_T) = \mathrm{VAR}(N_I + N_D) = n(\sigma_I^2 + \sigma_D^2 + 2C_{ID})$$

or $4\kappa \sim 2(2\alpha - 1)\bar{p}_I\bar{q}_I + 4\kappa(\bar{p}_I^2 + \bar{p}_D^2) + 2C_{ID}$. Finally,

$$COV(N_I N_T) = \bar{p}_I E(N_T^2) - E(N_I)E(N_T) = n\bar{p}_I\sigma_T^2 \sim \bar{p}_I 4\kappa n \ ,$$

and with (80), we obtain $COV(N_I N_Q) = -\frac{1}{2}COV(N_I N_T) \sim -\bar{p}_I 2\kappa n$.

To complete the proof, it suffices to check the asymptotic Gaussian property of N_I, N_D, N_Q. For N_Q, this follows from (80). For N_I, which is binomially distributed with parameter $\bar{p}_I$, we obtain, conditioning on N_T

$$\begin{aligned} E[e^{i\theta N_I/\sqrt{n}}] &= E\left\{[\bar{p}_I e^{i\theta/\sqrt{n}} + \bar{q}_I]^{N_T}\right\} = E\left\{[1 + \bar{p}_I i\,\frac{\theta}{\sqrt{n}} - \bar{p}_I\,\frac{\theta^2}{2n} + O(\frac{1}{n^{3/2}})]^{N_T}\right\} \\ &= E\left\{\exp N_T[i\bar{p}_I\frac{\theta}{\sqrt{n}} - \frac{\theta^2}{2n}\bar{p}_I\bar{q}_I + O(\frac{1}{n^{3/2}})]\right\} \end{aligned} \tag{82}$$

But, by (80),

$$E[e^{i\rho N_T}] = e^{in\mu_T\rho - \frac{1}{2}n\sigma_T^2\rho^2 + O(\rho^3\frac{n^{3/2}}{\sqrt{n}})}$$

hence by (82) we obtain

$$E[e^{i\theta N_I/\sqrt{n}}] = \exp\left\{in\mu_T\frac{\theta}{\sqrt{n}}\bar{p}_I - \frac{\theta^2}{2}(\bar{p}_I\bar{q}_I\mu_T + \sigma_T^2\bar{p}_I^2) + O(\frac{1}{\sqrt{n}})\right\}$$

which proves the asymptotic Gaussian property of N_I. ■

ACKNOWLEDGEMENT

We would like to thank Professors Luc Devroye, McGill University, and Michel Talagrand, Ohio State University and Paris VI, for discussions that led to our Theorem 2.3. We also appreciate detailed comments of one of the referees that helped us to avoid minor slips in some proofs.

References

[1] A. Apostolico and C. Guerra, The Longest Common Subsequence Problem Revisited, *Algorithmica*, 2, 315-336, 1987.

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

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

[4] 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.

[5] R. Arratia and M. Waterman, Critical Phenomena in Sequence Matching, *Annals of Probability*, 13, 1236-1249, 1985.

[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] R. Arratia and M. Waterman, A Phase Transition for the Score in Matching Random Sequences Allowing Deletions, *Annals of Applied Probability*, 1994.

[8] M. Atallah, P. Jacquet and W. Szpankowski, A Probabilistic Analysis of a Pattern Matching Problem, *Random Structures&Algorithms*, 4, 191-213, 1993.

[9] P. Billingsley, P., *Convergence of Probability Measures*, John Wiley & Sons, 1968

[10] J. Bucklew, *Large Deviation Techniques in Decision, Simulation, and Estimation*, John Wiley & Sons , 1990.

[11] A. Dembo and S. Karlin, Poisson Approximations for r-Scan Processes, *Annals of Applied Probability*, 2, 329-357, 1992.

[12] Z. Galil and K. Park, An Improved Algorithm for Approximate String Matching, *SIAM J. Computing*, 19, 989-999, 1990.

[13] W. Chang and J. Lampe, Theoretical and Empirical Comparisons of Approximate String Matching Algorithms, *proc. Combinatorial Pattern Matching*, 172-181, Tuscon 1992.

[14] V. Chvatal and D. Sankoff, Longest Common Subsequence of Two Random Sequences, *J. Appl. Prob.*, 12, 306-315, 1975.

[15] J. Griggs, P. Halton, and M. Waterman, Sequence Alignments with Matched Sections, *SIAM J. Alg. Disc. Meth.*, 7, 604-608, 1986.

[16] J. Griggs, P. Halton, A. Odlyzko and M. Waterman, On the Number of Alignments of k Sequences, *Graphs and Combinatorics*, 6, 133-146, 1990.

[17] W. Feller, *An Introduction to Probability Theory and its Applications*, Vol.I, John Wiley & Sons, 1970

[18] W. Feller *An Introduction to Probability Theory and its Applications*, Vol.II, John Wiley & Sons, 1971

[19] D.H. Greene and D.E. Knuth, *Mathematics for the Analysis of Algorithms*, Birkhauser, 1981

[20] D.L. Iglehart, Weak Convergence in Applied Probability, *Stoch. Proc. Appl.* 2, 211-241, 1974.

[21] S. Karlin and A. Dembo, Limit Distributions of Maximal Segmental Score Among Markov-Dependent Partial Sums, *Adv. Appl. Probab.*, 24, 113-140, 1992.

[22] S. Karlin and F. Ost, Counts of Long Aligned Word Matches Among Random Letter Sequences, *Adv. Appl. Prob.*, 19, 293-351 (1987).

[23] J.F.C. Kingman, *Subadditive processes*, in Ecole d'Eté de Probabilités de Saint-Flour V-1975, Lecture Notes in Mathematics, 539, Springer-Verlag, Berlin (1976).

[24] H.T. Laquer, Asymptotic Limits for a Two-Dimensional Recursion, *Stud. Appl. Math.*, 64, 271-277, 1981.

[25] T. Liggett *Interacting Particle Systems*, Springer-Verlag, New York 1985.

[26] G. Louchard, Random Walks, Gaussian Processes and List Structures, *Theor. Comp. Sci.*, 53, 99-124, 1987.

[27] G. Louchard, R. Schott and B. Randrianarimanana, Dynamic Algorithms in D.E. Knuth's Model : A Probabilistic Analysis, *Theor. Comp. Sci.*, 93, 201-225, 1992.

[28] G. Louchard and W. Szpankowski, A Probabilistic Analysis of a String Edit Problem, INRIA TR 1814, December 1992; revised Purdue University, CSD-TR-93-078, 1993.

[29] C. McDiarmid, On the Method of Bounded Differences, in *Surveys in Combinatorics*, J. Siemons (Ed.), vol 141, pp. 148-188, London Mathematical Society Lecture Notes Series, Cambridge University Press, 1989.

[30] E. Myeres, An $O(ND)$ Difference Algorithm and Its Variations, *Algorithmica*, 1, 251-266, 1986.

[31] C. Newman, Chain Lengths in Certain Random Directed Graphs, *Random Structures & Algorithms*, 3, 243-254, 1992.

[32] P. Pevzner and M. Waterman, Matrix Longest Common Subsequence Problem, Duality and Hilbert Bases, *proc. Combinatorial Pattern Matching*, 77-87, Tuscon 1992.

[33] D. Sankoff and J. Kruskal (Eds.), *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*, Addison-Wesley, Reading, Mass., 1983.

[34] E. Ukkonen, Finding Approximate Patterns in Strings, *J. Algorithms*, 1, 359-373, 1980.

[35] M. Waterman, L. Gordon and R. Arratia, Phase Transitions in sequence matches and nucleic acid structures, *Proc. Natl. Acad. Sci. USA*, 84, 1239-1242, 1987.

[36] M. Waterman, (Ed.) *Mathematical Methods for DNA Sequences*, CRC Press Inc., Boca Raton, (1991).