1988

(Probably) Optimal Solution to Some Problems Not Only On Graphs

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

Report Number:
88-780

TR # 780

(PROBABLY) OPTIMAL SOLUTION TO SOME PROBLEMS NOT ONLY ON GRAPHS

*Wojciech Szpankowski**
Department of Computer Science
Purdue University
West Lafayette, IN 47907

Abstract

A mathematical abstraction of the model studied in this paper can be formulated as follows: find the optimal value of $Z_{\max} = \max_{\alpha \in B_n} \{ \sum_{i \in S_n(\alpha)} w_i(\alpha) \}$, where n is an integer, B_n is the set of all *feasible solutions*, $S_n(\alpha)$ is a set of integers (objects), and $w_i(\alpha)$ is the *weight* assigned to the i-th object in the α-th feasible solution. Our interest lies in finding the asymptotically exact solution to this optimization problem in a probabilistic framework, that is, under assumption that the weights are random variables drawn independently and having identical distribution function. Such a formulation of the problem and the obtained solution, allows us to study in a uniform manner a large class of problems investigated vigorously in computer science over the last two decades. Among others we mention here: the assignment problem or perfect matching in bipartite graphs, the traveling salesman problem, the minimum spanning tree, the minimum weighted k-clique, the geometric location problem, the height of digital trees and so forth. Finally, we shall discuss algorithmic implications of the obtained solution.

1. MOTIVATION

Most algorithm designs are finalized to the optimization of the asymptotic worst-case performance. Insightful, elegant and generally useful constructions have been set up in this endeavor. Along these lines, however, the design of an algorithm is usually targeted at coping efficiently with unrealistic, even pathological inputs and the possibility is neglected that a simpler algorithm might perform just as well, or even better in practice.

This probabilistic approach to design algorithms was practically fulfilled a decade ago when it became clear that the prospects for showing the existence of polynomial time algorithms for NP-hard problems, were very dim. (Students of operations research, as opposed to those studying computer science, are convinced in the probabilistic heuristics, since in the very early

* This research w as supported in part by NSF grant NCR-8702115.

years of their study, they became familiar with the simplex method of linear programming which has exponential worst case behavior, but acceptable average case, i.e., practical complexity.) This fact, and apparently high success rate of heuristic approaches to solving certain difficult problems in practice, led Richard Karp [13] to undertake a more serious investigation of probabilistic approximation algorithms. The last few years witnessed an increasing interest in the probabilistic approach to the NP-hand problems [10,13,14,15,18,19,20,21,26].

Set aside the realm of approximation algorithms for NP-hand problems, then achieving a good average case performance is rarely the primary objective of algorithm design. This may surprise us, since algorithms that achieve this objective are also likely to be practically efficient. In assessing algorithmic performance, the average case analysis is often a more fruitful approach. By contrast, there are a number of algorithms and data structures for which the worst case analysis is unjustified or may lead to very expensive constructions. For example, in digital trees [2,15] with finite number of keys of (possible) unbounded length, the worst case analysis may lead for pathological input to unbounded search time. On the other hand, additional rebalancing constructions (e.g., AVL-tree [2,15]) applied to such digital trees, are very expensive operations. It turns out, however, that under mild assumptions, such trees are well balanced in practice [23], and therefore, the trees do not need to be restructured in order to keep them balanced.

Enlightened by these motivations, we undertake in this paper a study of a class of problems in a probabilistic framework. A general mathematical model of these problems can be formulated as follows: For every integer n, find the optimal value of $Z_{\max} = \max_{\alpha \in B_n} \{ \sum_{i \in S_n(\alpha)} w_i(\alpha) \}$ ($Z_{\min}$ respectively), where B_n is the set of all *feasible solutions*, $S_n(\alpha)$ is the set of all objects belonging to the α-th feasible solution, and $w_i(\alpha)$ is the weight assigned to the i-th object. For example, in the traveling salesman problem, B_n represents the set of all Hamiltonian paths in a graph with n vertices, $S_n(\alpha)$ is the set of edges which fall into the α-th Hamiltonian path, and

$w_i(\alpha)$ is the length of the i-th edge. In our probabilistic framework, we assume that the weights $w_i(\alpha)$ are random variables drawn independently with a common distribution function $F(\cdot)$. Our interest lies in finding the asymptotically exact values of all moments of Z_{max} and Z_{min}. In addition, we describe the asymptotic behavior of Z_{max} (Z_{min}) *in probability* and *almost surely* (with probability one) sense for a large class of distribution functions $F(\cdot)$. Finally, we apply these results to study heuristic algorithms for such problems as the traveling salesman problem [6,10,13,14,26], the assignment problem [6,9,19,25,26], the minimum spanning tree [4,5,12,15], the minimum weighted k-clique problem [19,4,5], the geometric location problem [21] and the height of digital trees [3,7,15,23,24].

In this paper, we would rather investigate a mathematical abstraction of a class of problems than a particular problem. Also, our solution, that is, asymptotically exact value of the objective function Z_{max} (Z_{min}), is general in the sense that a large class of distribution functions are considered. Some of the problems shown above have been investigated in the past [3,6,9,14,18,19,24,25,26], however, the approach undertaken in this paper is similar only to the work of Weide [26] and partially to Luker [19]. Nevertheless, Weide in his work has rather concentrated on (random) graphs, while we do not. We solve also, the open problem suggested by Weide, that is, we obtain asymptotically exact solutions in the cases the author of [26] provides only upper bounds. In addition, our techniques are completely different. Weide, as well as Luker [19] and others, in order to obtain their estimates, need to *know* the solution of the problem for *unweighted random graphs*. Our technique, which is very powerful and can be applied to many other problems (e.g., maximum queue length, etc.), avoids this requirement which, in fact, drastically limits application of Weide's results, even to graph problems.

This paper is organized as follows. In the next section, we rigorously formulate our problem, illustrate it in seven examples and present our main results. Section 3, provides proofs of these results, with some more interesting implications. Finally, Section 4 is more algorithmic

oriented and shows how our results can be applied to construct approximately sound algorithms to some interesting problems.

2. PROBLEM STATEMENT AND MAIN RESULTS

Let n be an integer (e.g., number of vertices in a graph, number of keys in a digital tree, etc.), which we further call the parameter of the problem. We are interested in the optimal value of $Z_{\max}$, ($Z_{\min}$) defined as

$$Z_{\max} = \max_{\alpha \in B_n} \sum_{i \in S_n(\alpha)} w_i(\alpha) \tag{2.1}$$

(for $Z_{\min}$ the operator max is replaced by minimum), where B_n is a set of all feasible solutions, $S_n(\alpha)$ is a countable set of objects which belong to the α-th feasible solution, and $w_i(\alpha)$ is the weight assigned to the i-th object in the α-th feasible solution. Throughout this paper, we adopt the following assumptions:

(A) The cardinality $|B_n|$ of B_n is fixed and equal to m. The cardinality $|S_n(\alpha)|$ of the set $S_n(\alpha)$ does *not* depend on $\alpha \in B_n$, and for all α it is equal to N, i.e., $|S_n(\alpha)| = N$.

(B) For all $\alpha \in B_n$ and $i \in S_n(\alpha)$ the weights $w_i(\alpha)$ are identically and independently distributed (i.i.d) random variables with common distribution function $F(\cdot)$.

The assumption (B) defines a probabilistic model of our problem (2.1), and therefore, we must investigate $Z_{\max}$ as a random variable. We shall ask for the asymptotic behavior of all moments of $Z_{\max}$ and $Z_{\min}$ as n becomes large. In addition, we present some results on the asymptotic behavior of $Z_{\max}$ that holds either *in probability* or *almost surely*, i.e., with probability one (see [8,19,22,26] for definitions). Before we plug onto the analysis, we discuss some important examples of our problem.

Example 2.1. *Linear assignment problem or perfect matching in bipartite graph*

Given an $n \times n$ matrix $\{a_{ij}\}_{i,j=1}^{n}$, the problem is to find a permutation α: $\{1, 2, \ldots, n\} \to \{1, 2, \ldots, n\}$ that maximizes or minimizes $\sum_{i=1}^{n} a_{i\alpha(i)}$. In our notations, B_n is a set of all permutations of $\{1, 2, \ldots, n\}$, $S_n(\alpha) = \{1, 2, \ldots, n\}$, and

$$Z_{\max} = \max_{\alpha \in B_n} \sum_{i=1}^{n} a_{i,\alpha(i)} \tag{2.2}$$

Note that $|B_n| = n$, $|S_n(\alpha)| = n$ and the weights $w_i(\alpha) = a_{i\alpha(i)}$. This problem is equivalent to the perfect matching in a bipartite graph [25].

Example 2.2. *Traveling salesman problem*

Let G_n be a graph with n vertices. We assign a (random) weight w_{ij} for every edge (i,j), $i,j = 1, 2, \ldots,$ belonging to the graph G_n. The traveling salesman problem is to find a path through all vertices with the minimum weight. Of course, this can be formulated as our problem (2.1) with B_n being the set of all Hamiltonian path and $S_n(\alpha)$ is a set of $n-1$ edges in the α-th Hamiltonian path, that is, $N = |S_n(\alpha)| = n - 1$. The cardinality of B_n depends on the structure of the graph, and general formula for $m = |B_n|$ can be found in [12]. For example, if G_n is a complete graph, then $|B_n| = (n-1)!$

Example 2.3. *The minimum spanning tree*

As in the previous example, the graph G_n with n vertices is given. We optimize $Z_{\min}$ with B_n interpreted as the set of all spanning trees, and $S_n(\alpha)$ with $|S_n(\alpha)| = N = n - 2$, as the set of edges belonging to the α-th spanning tree. The cardinality of B_n depends on the structure of G_n and a general formula on $|B_n|$ can be found in [12]. For example, in [12] we find that in a complete graph, there is $|B_n| = m = n^{n-2}$ rooted labeled trees.

Example 2.4. *The minimum weighted k-clique*

In a graph G_n with n vertices, we call a subgraph k-clique if it is spanned over k adjacent

vertices [4,5,19]. In addition, it is assumed that a weight w_{ij} is assigned to each edge (i,j) in G_n, $i,j = 1, 2, \ldots, n$. The objective function $Z_{\min}$ has the form of (2.1) with B_n being the set of all k-cliques and $S_n(\alpha)$ the set of edges belonging to the α-th k-clique. The cardinalities of $|B_n| = m$ and $|S_n(\alpha)| = N$, in general, depends on the structure of G_n, but for instance in the complete graph G_n, one immediately finds $m = \binom{n}{k}$ and $N = \binom{k}{2}$. Throughout the paper, we shall use the notation C_n^k for the Newton coefficient $\binom{n}{k}$, to simplify some of our formulas.

In the next two examples, either the weight function is not given explicitly or/and the distribution function $F(\cdot)$ of the weight must be computed from the model description.

Example 2.5. *Geometric location problem*

We present here only a one dimensional version of the problem. Let n points be randomly thrown into a line and let $k < n$. The problem is to select k points out of n, such that the distance from each point to the closest point in the set of k points achieves the minimum value. We can formulate this geometric location problem in terms of our notation, noting that the set of feasible solution is the set of all selection of k points out of n, that is, $|B_n| = m = C_n^k$. Of course $|S_n(\alpha)| = N = n$, but we must define the weight $w_i(\alpha)$. Let d_{ij} denote the distance between the i-th and the j-th point. Let $A_\alpha(k)$, $\alpha = 1, 2, \ldots, m$ contain the indices of the selected k points. Then

$$w_i(\alpha) = \min_{j \in A_\alpha(\alpha)} d_{ij}$$

The distribution of $w_i(\alpha)$ depends on the distribution function of d_{ij}. For example, if one assumes that n points are selected in such a way that the distance between two consecutive points is distributed according to a distribution function $F(\cdot)$, then $w_i(\alpha)$ is also $F(\cdot)$ distributed.

Example 2.6. *The height of digital trees (trie)*

In this example we deal with a digital data structure called a trie [2, 15], and our interest lies in computing the average value of the height. More precisely, let $X_1, X_2, \ldots, X_n$ be n strings of (possible) unbounded lengths formed by symbols from an alphabet Σ of cardinality V. For further simplifications, we assume a binary alphabet, e.g., $\Sigma = \{0,1\}$ and $V = 2$. (All results are trivially extended to V-ary alphabet.) A trie is built in a standard manner from the words (keys) $X_1, X_2, \ldots, X_n$ [2,24], that is, to insert a key, we split it into digits and 0 means "go left", and 1 means "go right" until an empty space is available for the insertion. All keys are stored in the so called external nodes and the access path from the root to an external node is a minimal prefix of the information contained in the external node. In [24], Szpankowski (see also [37]) has introduced the so called alignment or common operator, which is used to evaluate such tree parameters as height, depth, external path length, etc. The common (alignment) operator $C_{ij} = \mathrm{com}(X_i, X_j)$ is defined as the length of the longest string, that is, prefix of X_i and X_j. Thus, $C_{ij} = k$ iff X_i and X_j agree exactly on their first k positions, but differ on their $(k+1)$-st. In [24], it is shown that the height H_n of a trie built from $X_1, X_2, \ldots, X_n$ is given by

$$H_n = \max_{1 \le i < j \le n} \{C_{ij}\} + 1 \tag{2.2}$$

Formula (2.2) suggests that the problem can be reduced to our original formulation (2.1), if one defines $Z_{\max} = H_n$ and the weights as $\mathrm{com}\{X_i, X_j\}$. Moreover, the cardinality of the set of feasible solution B_n is obviously by (2.2) $m = n(n-1)/2$, while $|S_n(\alpha)| = N = 1$. The model will be fully described if one defines the probabilities framework, and one computes the distribution function for C_{ij}. We adopt the following two assumptions:

(i) The symbols of a word X_j, $j = 1, 2, \ldots, n$, are drawn independently from the alphabet $\Sigma = \{0,1\}$, and 0 occurs with probability p, while 1's with probability $q = 1 - p$.

(ii) The words $X_1, X_2, \ldots, X_n$ are statistically independent.

Assumptions (i)–(ii) defines the so called Bernoulli model. Then it is an easy exercise to see that

$$Pr\{C_{ij} = k\} = P^k(1-P) \quad k = 0, 1, \ldots, \tag{2.3}$$

where $P = p^2 + q^2$. Hence, the distribution function for the weights C_{ij} is $F(l) = Pr\{C_{ij} \le l\} = 1 - P^l$, and this completes the description of the model in terms of our original problem.

□

In some situations, our basic assumptions (A) and (B) are too restrictive. Therefore, we consider also two generalizations of our original problem, that is, we replace (A) and (B) by a more general assumptions:

(A′) The cardinality of $S_n(\alpha)$ depends on $\alpha \in B_n$, that is, $|S_n(\alpha)| = N_\alpha$.

(B′) The weights $w_i(\alpha)$ are dependent random variable with different distribution functions.

Finally, we generalize our original formulation (2.1). Let $f: R \to R$ be a function, then we define the objective function $Z^f_{\max}$ ($Z^f_{\min}$, respectively) as

$$Z^f_{\max} = \max_{\alpha \in B_n} \sum_{i \in S_n(\alpha)} f(w_i(\alpha)) \tag{2.4}$$

In practice our extended model finds many applications. We discuss below two of them.

Example 2.1a. *Assignment problem* – (continuation)

In Example 2.1, we have restricted ourselves to linear assignment problems. In general, we are interested in the extension of the basic model (2.2). For example, in the square assignment problem, the objective function is given by (i.e., $f(x) = x^2$)

$$Z^f_{\max} = \max_{\alpha \in B_n} \sum_{i=1}^{n} a^2_{i\alpha(i)} \tag{2.5}$$

Even more interesting extension can be obtained, if one assumes that the given matrix

$\{a_{ij}\}_{i,j=1}^{n}$ is *symmetric*, that is, in terms of perfect bipartite matching, we assume that the graph is undirected. Then assumption (B) does not hold any longer, since some of $a_{i\alpha(i)}$ might be dependent. For example, let $n = 2$ and $\alpha(1) = 2$, $\alpha(2) = 1$, hence $Z_{max} = a_{12} + a_{21} = 2a_{12}$, since $a_{12} = a_{21}$ by symmetry.

Example 2.7. *Suffix tree and the height of it*

Suffix tree is a digital tree (trie), as the one we discussed in Example 2.6, but the keys are very dependent. More precisely, let $X = x_1x_2x_3 \cdots$, be a string of (possible) unbounded length, and let $S_i = x_ix_{i+1} \cdots$ be the i-th *suffix* of X, $i = 1, 2, \ldots, n$. We store the first n suffixes of X in a trie in the same manner as discussed in Example 2.6. Such a digital tree is called suffix tree or position tree [2,3]. To analyze the height of it, we adopt our probabilistic framework from the Example 2.6 with the assumption (ii) obviously replaced by

(ii′) The words $S_1, S_2, \ldots, S_n$ are statistically dependent in the sense that they are consecutive suffixes of a given (random) word X.

Above in Example 2.6, we have argued that for any trie the height can be computed as in (2.2), that is, through the knowledge of (now called) self-alignments C_{ij} (i.e., $C_{ij} = k$ iff S_i and S_j agree exactly on k symbols, but differ on their $(k+1)$-st). Thus, we have reduced the problem of computing the height of (random) suffix tree to our original formulation, however, this time neither assumption (A) or (B) hold, but (A′) and (B′) are satisfied. Indeed, note that the weight, that is, the self-alignment C_{ij} depends on i and j, but fortunately in such a manner that the distribution of C_{ij} depends only on the difference $d = |j - i|$. Let us denote this random variable by C_d. For example, $C_{1,2}, C_{2,3}, \ldots, C_{n-1,n}$ have the same distribution as C_1 (i.e., $d = 1$). In order to complete the formulation of the model, we need to compute the distribution function of C_d, i.e., $F_d(k) = Pr\{C_d < k\}$. This is the most intricate computation we have discussed so far. Nevertheless, in [3] we have estimated the complement function of $F(k)$, that is, $R_d(k) = Pr\{C_d \geq k\}$. Let $k = dl + r$ where $l = 0, 1, 2, \ldots,$ and $r = 0, 1, \ldots, d-1$. Then

[3]

$$R_d(k) = (p^{l+2} + q^{l+2})^r (p^{l+1} + q^{l+1})^{d-r} \tag{2.6}$$

Note that for each $d = 1, 2, \ldots, n$ there are $n - d$ random variables C_d with the distribution function given by (2.6). We shall use this fact in Section 4 to evaluate the height of suffix trees.

□

In the rest of this section, we summarize our main results and discuss some of their implications. We present here only results for our basic model with assumptions (A) and (B). In Section 3 and 4, we discuss some extension of these results.

Under assumptions (A) and (B), the sum $\sum_{i \in S_n(\alpha)} w_i(\alpha)$ is a sum of $N = |S_n(\alpha)|$ i.i.d. random variables, each having distribution function $F(\cdot)$. We denote the distribution function of this sum by $F_N(\cdot)$, and from elementary probability, it is known that the density function $f_N(x) = F_N'(x)$ is a N-convolution of the densities $f(x) = F'(x)$ [8]. Let also $R(x)$ and $R_N(x)$ denote $1 - F(x)$ and $1 - F_N(x)$, respectively. We call $R(x)$ a reliability function. With this notation in mind, we can present our main result.

PROPOSITION. Let assumptions (A) and (B) hold, and the objective functions $Z_{\max}$ and $Z_{\min}$ is given by (2.1).

(i) If a_n and b_n are the smallest and the largest solutions of

$$mR_N(a_n) = 1, \qquad mF_N(b_n) = 1 \tag{2.7}$$

respectively, then for any distribution function $F(\cdot)$, the following bounds hold

$$EZ_{\max} \le a_n + m \int_{a_n}^{\infty} R_N(x)dx \tag{2.8a}$$

$$EZ_{\min} \ge b_n - m \int_{0}^{b_n} F_N(x)dx \tag{2.8b}$$

(ii) If the distribution function $F(\cdot)$ satisfies

$$F(x) < 1 \text{ for all } x \tag{2.9a}$$

$$\lim_{x \to \infty} \frac{1 - F(cx)}{1 - F(x)} = 0 \text{ for all } c > 1 \tag{2.9b}$$

then the following asymptotically exact results hold for all $r = 1, 2, \ldots,$

$$EZ_{\max}^r = a_n^r(1 + o(1)) \tag{2.10a}$$

$$EZ_{\min}^r = b_n^r(1 + o(1)) \tag{2.10b}$$

that is, $EZ_{\max}^r \sim a_n^r$ and $EZ_{\min}^r - b_n$ for the r-th moment of $Z_{\max}$ and $Z_{\min}$ respectively.

(iii) If hypothesis (2.9) of (ii) hold, then the following refinements are true

$$Z_{\max} \sim a_n, \; Z_{\min} \sim b_n \; \textit{in probability} \tag{2.11}$$

as n tends to infinity. In addition *almost surely* (i.e., with probability one)

$$\lim_{n \to \infty} \frac{Z_{\max}}{a_n} \leq 1, \; \lim_{n \to \infty} \frac{Z_{\min}}{b_n} \geq 1 \qquad w.p.1 \tag{2.12}$$

where *w.p.1* means with probability one.

□

The application of the Proposition crucially depends on satisfactory solutions of the following two problems:

(1) Explicit formula for the N-th convolution of the distribution function $F(x)$, that is, $F_N(x)$.

(2) Asymptotic solution to the (nonlinear) equations (2.7).

One way to get around these difficulties, is to consider special classes of distribution for which $F_N(x)$ can be computed. We investigate three kinds of distribution $F(x)$, namely

- gamma distribution *gamma* (β,λ) with the density function $f(x) = F'(x)$ given by [8,22]

$$f(x) = \frac{\lambda^{\beta}}{\Gamma(\beta)} x^{\beta-1} e^{-\lambda x} \quad x \geq 0 \tag{2.13}$$

- normal distribution $N(\mu,\sigma)$ with the density functions as below

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \tag{2.14}$$

- uniform distribution $U(0,1)$ over interval [0,1] for which the distribution function is

$$F(x) = x \qquad 0 \leq x \leq 1 \tag{2.15}$$

It is well known that the sum of N i.i.d. gamma distributions *gamma* (β,λ) and normal distributions $N(\mu,\sigma)$, are gamma $(N\beta,\gamma)$ and $N(N\mu,\sqrt{N}\sigma)$ respectively [8,22]. This implies that the distribution function $F_N(x)$ is

- for gamma distribution

$$F_N(x) = \frac{\gamma(N\beta,\lambda x)}{\Gamma(\beta)} \tag{2.16a}$$

where $\gamma(a,x)$ is the incomplete gamma function, that is [1]

$$\gamma(a,x) = \int_0^x t^{a-1} e^{-t} dt \tag{2.16b}$$

- for normal distribution

$$F_N(x) = \Phi\left[\frac{x - N\mu}{\sqrt{N}\sigma}\right] \tag{2.17a}$$

where $\Phi(x)$ is the error function defined as [1]

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-y2/2} dy \tag{2.17b}$$

- for uniform distribution, see Feller [8] where it is proved that

$$F_N(x) = \frac{1}{N!} \sum_{k=0}^{N} (-1)^k \binom{N}{k} (x-k)_+^n$$
$$\text{where } x_+ \stackrel{def}{=} \max\{0,x\}. \tag{2.18}$$

Thus, with (2.16)–(2.18) the Proposition can be directly applied, assuming one solves nonlinear equations (2.7). Note that the gamma distribution and the normal distribution satisfies conditions (2.9) of Proposition (ii), while the uniform distribution not, hence Proposition (i) must be applied in the latter case. Some particular, but useful solutions of (2.7), are summarized in the following Corollary, which will be formally proved in the next section.

COROLLARY. If the weights in our basic problem (2.1) are distributed according to:

(i) *gamma* (β,λ) with $|B_n| = m = n!$ and $N = n$, then

$$EZ_{\max} \sim \frac{n}{\lambda} \log n + \frac{n\beta}{\lambda} \log\log n + O(n) \tag{2.19}$$

$$EZ_{\min} \sim \frac{n^{1-1/\beta}}{\lambda} + o(n^{1-1/\beta}) \tag{2.20}$$

where $\log(\cdot)$ denotes the natural logarithm, and $Z_{\max}$, $(Z_{\min}) \sim f(n)$ means *asymptotically equal* in the sense $Z_{\max} = f(n)\,(1 + o(1))$.

(ii) normal distributions $N(\mu,\sigma)$, then

$$EZ_{\max} \sim N\mu + \sigma\sqrt{N}\,\sqrt{2\log m - \log\log m + O(1)} \tag{2.21}$$

$$EZ_{\min} \sim N\mu + \sigma\sqrt{N}\,\sqrt{2\log m - \log\log m + O(1)} \tag{2.22}$$

(iii) uniform distribution $U(0,1)$, with $|B_n| = m = n!$

$$EZ_{\max} \le n - \frac{n}{n+1} \tag{2.23}$$

$$EZ_{\min} \ge 1 - \frac{1}{n+1} \tag{2.24}$$

□

The corollary solves, in some sense, the two difficulties (1) and (2) mentioned just after the Proposition. Nevertheless, in the corollary, we restrict our interest only to a special class of distributions for which $F_N(x)$ is known. However, there is a possibility to obtain the leading factor in the asymptotics of $EZ_{\max}$ and $EZ_{\min}$ (in fact, any moment of $Z_{\max}$ and $Z_{\min}$) without knowing $F_N(x)$. Indeed, let μ and σ^2 be the average and the variance of the weight $w_i(\alpha)$, that is, $\mu = Ew_i(\alpha)$ and $\sigma^2 = \text{var}\, w_i(\alpha)$. Rewriting (2.1) in the following form

$$Z_{\max} = N\mu + \sigma\sqrt{N} \max_{\alpha \in B_n} \left\{ \frac{\sum_{i \in S_n(\alpha)} w_i(\alpha) - N\mu}{\sigma\sqrt{N}} \right\} \tag{2.25}$$

one shows immediately that the expression in the square brackets tends to the normal distribution because of the famous *central limit theorem* [8]. It is very tempting to draw quick conclusions. Nevertheless, we not that solving for a_n and b_n in (2.7), we *must* know the rate (error) in which our expression tends to the normal distribution. Then depending on the values of n, we *may* or *may not* use the central limit theorem (the error might be comparable with m^{-1} and the solution of (2.7) might give an incorrect answer). Details are provided in the next section, while here we present our main conclusion in the form of the following.

THEOREM (Normal approximation). If $m = O(N^p)$ where p is a fixed constant, and the distribution function $F(\cdot)$ of the weights satisfies (2.9), then for $N \to \infty$ as $n \to \infty$

$$EZ_{\max} \sim N\mu + \sigma\sqrt{2N_p \log N} + o(\sqrt{N \log N}) \tag{2.26a}$$

$$EZ_{\min} \sim N\mu - \sigma\sqrt{2N_p \log N} + o(\sqrt{N \log N} \tag{2.26b}$$

□

Note for example that (2.26) does not hold for gamma distribution with $m = n!$ (see (2.19), (2.20)), simply because m is not of a polynomial order of n.

Finally, we briefly comment here on the application of our result to the design of approximate algorithms. A more detailed discussion of this issue is contained in the last section of the

paper. Our Proposition and the Corollary, as well as the Theorem, provide under some mild condition of the distribution function $F(\cdot)$, the leading factor in the asymptotic approximation of all moments of $Z_{\max}$, as well as convergence of $Z_{\max}$ *in probability* and *with probability one.* This tells us how to construct a sound approximate algorithm for some NP-hard problems in the sense that the relative error between a heuristic and the optimal algorithms tends to zero as the size of the problem becomes large. For example, postulate for a moment that our Propositions show that the true optimal solution to an NP-hard problem (e.g., traveling salesman problem), is $Z_{\max} = n \log n\,(1 + o(1))$. This implies that any heuristic, no matter how simple, which solves the problem and gives the value of $Z_{\max}$ asymptotically equal to $n \log n$ is a *sound* approximation, in the sense that the relative errors tends to zero as the size of the problem tends to infinity [26]. In most cases, such heuristics are easy to construct and they naturally arise from the problem description. Moreover, based on our Theorem, we can suggest general rules on how to construct good heuristics. We delay detailed discussions to the last section.

3. ANALYSIS AND MORE RESULTS

In this section, we prove our Proposition, Corollary and Theorem, providing in addition some more results. Our main result in this section is Lemma 1, which is a "locomotive" for all of our results, and has a flavor of methodological approach useful in many other problems dealing with maximum of dependent random variables.

3.1 How to prove our proposition

Our main result, i.e., our Proposition, is based on the following Lemma concerning maximum of dependent random variables. Let $Y_1, Y_2, \ldots, Y_m$ be a sequence of random variables with distribution functions $G_1(y)$, $G_2(y)$, . . . , $G_m(y)$ respectively. Let also $R_i(y) = 1 - G_i(y) = Pr\{Y_i \geq y\}$ be the complement function of $G_i(y)$ (i.e., $R_i(y)$ is called reliability function). We are interested in the maximum and minimum of $Y_1, Y_2, \ldots, Y_m$, that is,

$$\overline{M}_m = \max_{1 \le i \le m} \{Y_i\}; \qquad \underline{M}_m = \max_{1 \le i \le m} \{Y_i\}$$

The following Lemma extends slightly, the not too well known results of Lai and Robbins [16,17].

Lemma 1. (i) If a_m and b_m are solutions of the equations

$$\sum_{k=1}^{m} R_k(a_m) = 1 \quad \sum_{k=1}^{m} G_k(b_m) = 1 \tag{3.1}$$

respectively, then the following inequalities hold

$$E\overline{M}_m \le a_n + \sum_{k=1}^{m} \int_{a_n}^{\infty} R_k(x)dx \tag{3.2a}$$

$$E\underline{M}_m \ge b_m - \sum_{k=1}^{m} \int_{-\infty}^{b_n} G_k(x)dx \tag{3.2b}$$

(ii) If $Y_1, Y_2, \ldots, Y_m$ are identically distributed with the common distribution function $G(y)$ that satisfies

$$F(y) < 1 \text{ for all } y < \infty \tag{3.3a}$$

$$\lim_{y \to \infty} \frac{1 - G(cy)}{1 - G(cy)} = 0 \text{ for all } c > 1 \tag{3.3b}$$

then for any $r = 0, 1, \ldots$

$$\lim_{m \to \infty} \frac{E\overline{M}_m^r}{a_m^r} = \lim_{m \to \infty} \frac{E\underline{M}_m^r}{b_m^r} = 1 \tag{3.4}$$

where a_m and b_m are the smallest and the largest roots of

$$mR(a_m) = 1, \qquad mG(b_m) = 1 \tag{3.5}$$

(iii) Let hypotheses of (ii) hold. Then, in addition,

$$\lim_{m \to \infty} \frac{\overline{M}_m}{a_m} = \lim_{m \to \infty} \frac{\underline{M}_m}{b_m} = 1 \ \textit{in probability} \tag{3.6}$$

and

$$\lim_{m \to \infty} sup \ \frac{\overline{M}_m}{a_n} \le 1, \quad \lim_{m \to \infty} inf \ \frac{\underline{M}_n}{b_m} \ge 1 \qquad w.p.\,1. \tag{3.7}$$

Proof. We restrict the analysis to $\overline{M}_m$, since $\underline{M}_m = \max_{1 \le i \le n} \{-Y_i\}$. For part (i) note that for any a_m

$$\overline{M}_m \le a_m + \sum_{k=1}^{m} [Y_k - a_m]_+ \tag{3.8}$$

where $x_+ = \max\{x, 0\}$. Computing the average of (3.8) and minimizing the RHS of (3.8) with respect to a_m, one finds conditions (3.1) (for details see [24,3]). Finally, note that $M_m^r = \max_{1 \le i \le m} \{Y_i^r\}$. Part (ii) and the second part of (iii), are proved in [16,17], so we concentrate here on the first part of (iii). Note, that by (3.4) we have $EM_n^2 = a_m^2 \ (1 + o(1))$ and $EM_m = a_m(1 + o(1))$, hence also $\text{var}\, M_m = a_n^2 \cdot o(1)$. Then, by Chebyshev inequality [8,22]

$$Pr\left\{\left|\frac{M_m}{EM_m} - 1\right| > \varepsilon\right\} = \frac{\text{var}\, M_m}{\varepsilon^2 (EM_m)^2} = o(1)$$

so (3.6) follows.

□

In order to apply Lemma 1 to our basic problem (2.1), note that $Z_{\max}$, ($Z_{\min}$) can be equivalently represented as $Z_{\max} = \max_{1 \le \alpha \le m} \{Y_\alpha\}$ where $Y_\alpha = \sum_{i \in S_n(\alpha)} w_i(\alpha)$. But, under assumptions (A) and (B), the distribution function $G(x)$ of Y_α (more precisely, the density function $g_N(x) = G(x)$) is the N-convolution of densities of the weights $f(x) = F'(x)$. We denote this distribution function by $F_N(x)$. Then, applying directly Lemma 1 to $Z_{\max} = \max_{1 \le \alpha \le m} \{Y_\alpha\}$ with the distribution function $F_N(x)$, one proves Proposition (i). To prove Proposition (ii) and (iii), we need to show conditions (2.9) are equivalent to (3.3). We prove

Lemma 2. If $F(x)$ satisfies (2a) and (2b), then $F_N(x)$ satisfies

$$F_N(x) < 1 \ \text{ for all } \ x \tag{3.9a}$$

$$\lim_{y \to \infty} \frac{1 - F_N(cx)}{1 - F(x)} = 0 \text{ for all } c > 1 \tag{3.9b}$$

Proof. The condition (3.9a) trivially follows from (2.9a). To prove (3.9b), note that

$$1 - F_N(cx) = Pr\{Y_1 + Y_2 + \cdots + Y_N > cx\} \le N[1 - F(cx)]$$

$$1 - F_N(x) = Pr\{Y_1 + Y_2 + \cdots + Y_N > x\} \ge Pr\{Y_1 > x\} = 1 - F(x)$$

therefore (3.9b) follows immediately from (2.9b) and the above.

□

Using Lemma 2 and Lemma 1 (ii), (iii), we prove our Proposition part (ii) and (iii).

3.2 The corollary is easy to prove

In this subsection, we prove three statements (i)–(iii) from our Corollary. We start with the gamma distribution (2.13) with parameters β and λ. Assuming $|S_n(\alpha)| = N = n$, we note that the sum of n i.i.d. gamma distribution *gamma* (β,λ) is *gamma* $(n\beta,\lambda)$. Then simple arguments lead to the following formulas on the distribution function $F_n(x)$ and the reliability function $R_n(x)$ of *gamma* $(n\lambda,x)$

$$F_n(x) = \frac{\gamma(n\beta,\lambda x)}{\Gamma(\beta)}, \qquad R_n(x) = \frac{\Gamma(n\beta,\lambda x)}{\Gamma(\beta)} \tag{3.10}$$

where the incomplete gamma functions $\gamma(a,x)$ and $\Gamma(a,x)$ are defined as [1,11]:

$$\gamma(a,x) = \int_0^x e^{-t} t^{a-1} dt, \qquad \Gamma(ax) = \int_x^\infty e^{-t} t^{a-1} dt \tag{3.11}$$

The purpose of our analysis is to derive asymptotic approximation for solutions a_n and b_n of equations (2.7). These solutions strongly depend on the value of m. From the application viewpoint, the case $m = n!$ is the most interesting, and we restrict our investigation to that case. Also, for simplicity of algebraic manipulation,, we assume $\beta = 1$. At first, we consider Z_{max}, that is, we search for solution a_n of $n!R_n(a_n) = 1$. It is known that for $n > 1$ and $x > n - 1$ [1]

$$e^{-\lambda x}(\lambda x)^{n-1} \le \Gamma(n,\lambda x) \le \frac{e^{-\lambda x}(\lambda x)^n}{(\lambda x - n + 1)}$$

This and a rough estimation of a_n (i.e., $a_n = O(n \log n)$), suggest to approximate $\Gamma(n, \lambda x)$ by the asymptotic formula [11] $\Gamma(n, \lambda x) \sim e^{-\lambda x}(\lambda x)^n$ for $x \to \infty$. Then, the problem lies in solving the following equation ($\lambda = 1$)

$$e^{-a_n}(a_n)^{n-1} = 1$$

or equivalently

$$a_n - (n-1)\log a_n - \log n = 0 \tag{3.12}$$

for large n. Let the LHS of (3.12) be denoted as $f(a_n)$. We find such $\underline{a}_n$ and $\bar{a}_n$ that $\underline{a}_n \le a_n \le \bar{a}_n$, that is, $f(\underline{a}_n) > 0$ and $f(\bar{a}_n) < 0$. Let

$$\underline{a}_n = n \log n + n \log\log n$$

Then

$$f(\underline{a}_n) = n \log\log n - (n-1) \log\log n \log n < 0$$

for large n. On the other hand, for any $\varepsilon > 0$ define

$$\bar{a}_n = n \log n + n \log\log n^{1+\varepsilon}$$

Note that

$$f(\bar{a}_n) = n \log\log n^{1+\varepsilon} - (n-1) \log\log n \log n^{1+\varepsilon} > 0$$

for sufficiently large n. Hence we prove that $a_n = n \log n + n \log\log n + O(n)$. Dividing this by λ we finally show (2.19) in Corollary (i). The proof for $b_n \sim EZ_{\min}$ goes the same way, except that $\Gamma(n,x)$ is replaced by $\gamma(n,x)$ and the following asymptotic approximation $\gamma(n,x) \sim \frac{x^n e^{-x}}{n}$ is used [11], since we know that b_n is bounded for $\beta = 1$. Details are omitted and left for interesting readers.

In the proof of (2.21) and (2.22) in Corollary (ii) we use the following representation of our original problem

$$Z_{\max} = N\mu + \sigma\sqrt{N}\left\{\frac{\sum_{i \in S_n(\alpha)} w_i(\alpha) = N\mu}{\sigma\sqrt{N}}\right\} \tag{3.13}$$

Then, the expression in the parentheses is normalized normal distribution with distribution function $\Phi(x)$ as defined in (2.17b). Using the following inequality [8]

$$\frac{1}{\sqrt{2\pi}} e^{-x^2/2} \left(\frac{1}{x} - \frac{1}{x^3} \right) \leq 1 - \Phi(x) \leq \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \tag{3.14}$$

or equivalently $\Phi(x) \sim e^{-x^2/2}/(x\sqrt{2\pi})$ for $x \to \infty$, and applying the same line of arguments as above, we finally prove Corollary (ii). The proof of the last part of the Corollary for the uniform distribution of the weights is rather simple. We just note that for b_n we need to solve $x^n = 1$, so $b_n = 1$, while for $a_n = n - 1$ solves (2.7). Since the uniform distribution $U(0,1)$ does not satisfy (2.9a), we cannot apply Lemma 1 (i), and after computing the integrals, the proof of the Corollary is done.

3.3 Be careful with the normal approximation

It is tempting to apply the central limit theorem and approximate the expression in parentheses in (3.13) by the normalized normal distribution $N(0,1)$. However, one must be very careful with such an approximation and this subsection shows how to cope with the problem, that is, how to prove our theorem.

Let for the purpose of this subsection, $F_N(x)$ denote the distribution function of $\left[\sum_{i \in S_n(\alpha)} w_i(\alpha) - N\mu \right] / \sigma\sqrt{N}$ where μ and σ^2 are the average and the variance of the weights $w_i(\alpha)$ respectively. By the central limit theorem, we know that $\lim_{N \to \infty} F_N(x) = \Phi(x)$ where $\Phi(x)$ is the error function given in (2.17b). So $F_N(x) = \Phi(x) + e(x)$ where $e(x)$ is the error function, and the value of $e(x)$ is crucial for the solution of $mR_N(x) = 1$ and $mF_N(x) = 1$ for a_n and b_n in our Proposition. In other words, to obtain sound approximation of a_n and b_n using the central limit theorem approach, the following condition $e(x) = o(m^{-1})$ must hold. Indeed, the equation $mR_N(x) = 1$ leads to $\Phi(x) + e(x) = \frac{1}{m}$, and the values of $e(x)$ can be omitted if $e(x) << m^{-1}$.

From Feller [8], we know that

$$F_N(x) = \Phi(x) + \phi(x) \sum_{k=3}^{r} N^{-1/2k+1} P_k(x) + o(N^{-1/2r+1}) \tag{3.15}$$

where $\phi(x) = e^{-x^2/2}/\sqrt{2\pi}$ and $P_k(x)$ is a polynomial of degree k, dependent only of the moments of the weights $w_i(\alpha)$, but *not* on N and r, where r is any integer (the larger r is the better the approximation is). For practical purposes, we can approximate $1 - \Phi(x) \sim e^{-x^2/2}(x\sqrt{2\pi})$. Note also, that the polynomial $P_k(x)$ does not change significantly the asymptotic solution of (2.7) (since exponential "swallows" polynomials). Therefore, selecting r such that $N^{-1/2r+1} = o(m^{-1})$ is sufficient to obtain exact (with respect to the leading factor), asymptotics for a_n and b_n. For instance, $m = O(N^p)$, p is a constant, satisfies this condition, hence with the help of Corollary (ii), we prove the theorem.

4. APPLICATIONS – ALGORITHMIC APPROACH

A non-mathematical reader of this paper may ask what are practical implications of the main result. In this section, we first present some general conclusions of our work, and then discuss in details, the seven examples from Section 2.

First of all, we note that knowing exact asymptotic expansion for $Z_{\max}$, ($Z_{\min}$) in the average, the probability or the almost surely sense, helps to design heuristic algorithms in the sense that the relative error tends to zero, as the size of the problem becomes large. Indeed, let us postulate that using our Proposition, one proves that $EZ_{\max}^r = a_n^r\ (1 + o(1))$. This is the *exact optimal* value of the objective function for a large value of n. Furthermore, we assume that to achieve this optimal value for all inputs is too expensive (e.g., NP-hard problem), so one constructs algorithms which has the same leading factor in the asymptotic approximation. Mathematically speaking, this means that the approximate new objective function $Y_{\max}$ satisfies asymptotically $EY_{\max}^r = a_n^r\ (1 + o(1)))$ (for simplicity, it is assumed that a_n represents only leading factors in the asymptotic of $EZ_{\max}^r$). Then, the following "error lemma" justifies our

previous claim.

Lemma 3. The relative error $\varepsilon_n^r = (EZ_{\max}^r - EY_{\max}^r)/EZ_{\max}^r$ of the approximate algorithm represented by $Y_{\max}$ is small for large n, that is

$$\lim_{n \to \infty} \varepsilon_n^r = 0 \tag{4.1}$$

Proof. Since $EZ_{\max}^r = a_n^r\,(1 + o(1))$ and $EY_{\max}^r = a_n^r\,(1 + o(1))$, then

$$\varepsilon_n^r = \frac{o(1)}{1 + o(1)} = o(1)$$

which proves the Lemma.

□

These arguments can easily be extended to prove (4.1), in the sense of convergence *in probability* and *almost sure* convergence. This follows immediately from Proposition (ii)–(iii) (see also [25]).

Lemma 3 justifies approximate algorithms, but it does not tell how to construct sound approximations. This, however, can be done by applying our Proposition, in particular Part (ii) and (iii). The Proposition solves half of the problem, namely, tells us what is optimal value of $EZ_{\max}$, hence it might suggest potential heuristics. A general result, how to construct "optimal" approximation in the sense of Lemma 3, can be obtained from Corollary (i) and (ii). Indeed, for example for *gamma* $(1,\lambda)$ (i.e., exponential) distribution of weights with $m = n!$ (2.19) and (2.20), suggest that

$$E \max_{\alpha \in B_n} \sum_{i \in S_n(\alpha)} w_i(\alpha) \sim \sum_{i \in S_n(\alpha)} E \max_{1 \le \alpha \le n} w_i(\alpha) \tag{4.2}$$

since $E \max_{1 \le i \le n} w_i(\alpha) \sim \frac{1}{\lambda} \log n$ for $w_i(\alpha)$ exponentially distributed [8], [22]. The same holds for normal distribution if one assumes, again, $m = n!$. Note also that (4.2) does *not* hold for $m = O(n^p)$, p is a constant as proved in the theorem. Therefore, property (4.2) may be satisfied only for problems with nonpolynomial cardinality of the set B_n with respect to n.

If property (4.2) holds, then one can build a general heuristic algorithm, which is optimal in the sense of Lemma 3 and which runs no faster than $O(Nn)$. The idea of the algorithm is to find first the maximum (minimum) in the sets $\{w_i(1), \ldots, w_i(n)\}$ for each i, and then to sum it up. This can be written in a pseudo-algol language as follows. Let L be a list of objects in the (approximately) optimal solution of a problem satisfying (4.2). Then, the greedy algorithm is

```
begin
        L = empty list;
for i = 1 until N do
        begin
            find α* which maximizes (minimizes) w_i(1), ..., w_i(n);
            append α* to L;
        end;
end;
```

Figure 1

This program presents greedy algorithms for a class of problems, satisfying property (4.2). The question is how to know whether a problem satisfies (4.2) or not. A partial answer can be extracted from our Proposition, Corollary and Theorem. We may conclude that property (4.2) (or its slight modification), holds if the distribution function satisfies conditions (2.9), $N = O(n)$, and m increases faster than polynomial with respect to n (see Theorem). Nevertheless, a full answer to that question seems to be open.

Finally, we discuss in some detail, the seven motivating examples from Section 2, and provide more information on how to construct good heuristics for these algorithms.

Example 4.1. *Assignment problem revisited*

In that case $|B_n| = n!$ and $N = n$, hence for distribution function $F(\cdot)$ satisfying (2.9) (e.g., exponential and normal distributions), we can apply our greedy algorithm, since property (4.2) holds, and the optimal values of $Z_{\max}$ and $Z_{\min}$ are given in the Corollary (see also [6,19,26]). In particular, the greedy algorithm works as follows: (1) take the minimum element

from the first column and delete the row in which you found the element, (2) consider the second column (only $n-1$ elements left), find the minimum element and strike out the row with the element, (3) and so on. Note, however, that the greedy algorithm does not work for distributions, which do not satisfy conditions (2.9), e.g., uniform distribution. A greedy version for such distribution was suggested by Walkup [25], who proposed to consider not the minimum element in each column, but the $c > 1$ smallest ones in each column, and he proved that a greedy algorithm gives approximation for which the errors lends to zero for large n. The question is whether this argument can be extended to build a general heuristic, as in Figure 1 for distribution functions not satisfying condition (2.9).

Example 4.2. *Traveling salesman problem revisited*

We can apply our results to any graph for which we are able to estimate the number of Hamiltonian paths. In particular, if the graph is complete, then $|B_n| = (n-1)!$ and $N = n$, hence Corollary can immediately apply. Also, for exponential and normal distributions, we can use our general scheme in Figure 1 to construct greedy algorithms.

Example 4.3. *The minimum spanning tree revisited*

Actually, everything we said in Example 4.2, can be applied to this case. In particular, for complete graph $N = m - 2$, $m = n^{n-2}$, and the leading factor in formulas (2.19)–(2.22) of the Corollary is unchanged. A greedy algorithm based on the scheme in Figure 1 can be constructed in a similar manner.

Example 4.4. *The minimum weighted k-clique revisited*

This is a slightly different problem and the Corollary cannot be directly applied, however, we may use our Proposition. In Example 2.4, it was shown that for this problem the cardinality of B_n and $S_n(\alpha)$ are $m = C_n^k$ and $N = C_k^2$ respectively. Note that N does *not* tend to infinitely with n, hence the Theorem cannot be applied. However, assuming that weights are normally

distributed $N(\mu,\sigma)$ with parameter μ and σ, the arguments used in the proof of Corollary (ii) can be easily extended to show that

$$EZ_{\max} \sim \mu C_k^2 + \sigma\sqrt{2C_k^2 \log C_n^k} \sim \mu C_k^2 + \sigma k\sqrt{(k-1)\log n} \tag{4.3a}$$

$$EZ_{\min} \sim \mu C_k^2 - \sigma k\sqrt{(k-1)\log n} \tag{4.3b}$$

where in the RHS of (4.3), we use the approximation $C_n^k \sim n^k/k!$ for large n and bounded k. A little more intricate situation arises in the case of exponential distribution of weights. In general, the following two equations must be solved (see (3.10), (3.11))

$$C_n^k \Gamma(C_k^2, a_n) = \Gamma(C_k^2), \qquad C_n^k \gamma(C_k^2, b_n) = \Gamma(C_k^2) \tag{4.4}$$

to obtain asymptotics $EZ_{\max} \sim a_n$ and $EZ_{\min} \sim b_n$. Nevertheless, using our estimate of the incomplete gamma functions, we may prove using the same arguments as in Section 3.2, the following approximations

$$EZ_{\max} \sim (n-k)\log(n-k) \tag{4.5a}$$

$$EZ_{\min} \sim n^{\frac{(k-1)}{2}} [k!\Gamma(C_k^2+1)]^{1/C_k^2} \tag{4.5b}$$

Formulas (4.3) and (4.5) suggest that a greedy algorithm for these distributions can be constructed. It imitates the algorithm in Figure 1 with some necessary modification. In the case of normal distributions, the details of such an algorithm can be also found in [19].

Example 4.5. *Geometric location problem revisited*

In that case $m = C_n^k$ and $N = n$, where n is the number of points and k is the set of the selected points, $k < n$. The weight is understood to be the distance between two closest points, and we assume distribution functions of such weights are given. For fixed k and bounded away from n (i.e., $k << n$), $m \sim n^k/k!$, therefore, we can apply our Theorem to show that for distributions $F(\cdot)$ satisfying (2.9) the following asymptotics hold

$$EZ_{\max} \sim n\mu + \sigma\sqrt{2nk\log n} \tag{4.6a}$$

$$EZ_{\min} \sim n\mu - \sigma\sqrt{2nk \log n} \tag{4.6b}$$

where μ and σ^2 are the mean and the variance of the distances distributed according to $F(\cdot)$. Note that in that case, the greedy algorithm described in Figure 1 cannot be applied, since m is polynomially related to n.

Example 4.6. *Digital trie revisited*

In Section 2, we show that the height of a trie can formulate in terms of our problem (cf. (2.2)). In this case, the weights are interpreted as the alignment C_{ij} defined as the common number of digits (i.e., prefix) of the i-th and the j-th key. By assumptions (i)–(ii) from Example 2.6, we have argued that the distribution of C_{ij} is geometric with parameter $P = p^2 + q^2$. Noting that $R(k) = Pr\{C_{ij} \geq k\} = P^k$ and using Lemma 1 (ii), we immediately prove that

$$EH_n = \frac{2}{\log P^{-1}} \log n(1 + o(1)) + 1 \tag{4.7}$$

i.e., $EH_n \sim -2\log_P n$. For example, for binary symmetric case $EH_n \sim 2\log_2 n$. Furthermore, using Proposition (iii), we show that $EH_n \sim -2\log_p n$ *in probability*. These results, together with those obtained in [23], suggest that digital trees are well balanced (in particular, for the symmetric case), and they do not need to be restructured to keep them balanced.

Example 4.7. *Suffix tree revisited*

The suffix tree and computation of the height of it through the approach taken in our paper, is described in Example 2.7. In particular, we note that computation of the average height EH_n of the tree can be reduced to our problem with $m = n^2$ and $N = 1$, but this time the n^2 variables are not identically distributed. Therefore, Lemma 1 (i) has to be used, and by (2.6), (3.1), we seek for the solution of a_m of

$$\sum_{d=1}^{n} (n - d)R_d(a_n) = 1 \tag{4.8}$$

where $R_d(x)$ is given in (2.6). Formula (4.8) follows from the fact that there are $(n - 1)$

alignment variables C_1, $(n-2)$ alignment variables $C_2, \ldots,$ and one variable C_{n-1}. The solution of (4.8) can be upper bounded by a solution of the following simpler equation (for details see [3])

$$m(p^{\bar{a}_n+1} + q^{\bar{a}_n+1}) = 1 \tag{4.9}$$

where $m = n^2$ and $a_n \le \bar{a}_n$. The asymptotic solution to (4.9) can be easily obtained and one proves

$$\bar{a}_n = 2\log_{p_{\max}} n^{-1} + 1 + o(1) \tag{4.10}$$

where $p_{\max} = \max\{p,q\}$ (in general, $p_{\max} = \max\{p_1, p_2, \ldots, p_V\}$). Finally, to complete our analysis, we need to evaluate the second term in (3.2a), that is, in our case $\sum_{j=a_n}^{\infty} \sum_{d=1}^{n} (n-d)R_d(j)$. Using (4.8) and the bound (see [3] for details).

$$R_d(k) \le (p^{f+1} + q^{f+1})^d$$

where $f = \lfloor \frac{k}{d} \rfloor$ and $\lfloor \cdot \rfloor$ is the floor operation, we prove

$$\sum_{j=a_n}^{\infty} \sum_{d=1}^{n} (n-d)R_d(j) = \sum_{k=0}^{\infty} \sum_{d=1}^{n} (n-d)R_d(a_n+k) = O\left[\sum_{d=1}^{\infty}(n-d)R_d(n) \sum_{k=0}^{\infty} p_{\max}^k\right] = O(1)$$

Hence, by the above and (4.10) we finally obtain

$$EH_n \le \frac{2}{\log_{p_{\max}^{-1}}} \log n + c \tag{4.11}$$

where c is a constant. Note, that in the symmetric case, all self-alignments C_{ij} are identically distributed (e.g., set $p = q$ in (2.6)), and then by Lemma 1 (ii), we obtain stronger results

$$EH_n \sim 2\log_V n \tag{4.12}$$

The consequences of these results, are discussed in details in [3]. Here we only point out that (4.11) suggests a direct, natural construction of a suffix tree (that is, by consecutive insertion of suffixes) takes $O(n \log n)$ time in average, while rather sophisticated methods takes $O(n)$ time, however, the latter uses much more complicated data structure. On the other hand, using this

direct construction, one can prove that computing the full statistics without overlap of all substrings of a word takes $O(n \log n)$ expected time, while a more sophisticated method oriented on the worst case analysis, takes $O(n \log^2 n)$, and so on.

References

[1] Abramowitz, M. and Stegun, I., *Handbook of mathematical functions*, Dover, New York (1964).

[2] Aho, A., Hopcroft, J. and Ullman, J., *The design and analysis of computer algorithms*, Addison-Wesley, Reading (1974).

[3] Apostolico, A. and Szpankowski, W., Self-alignments in words and their applications, Purdue University CSD TR-732, 1988 (submitted to a journal) .

[4] Berge, C., *Graphs and hypergraphs*, North-Holland, Amsterdam (1973).

[5] Bollobas, B., *External graph theory*, Academic Press, London (1978).

[6] Borovkov, A., A probabilistic formulation of two economic problems, *Soviet Mathematics*, 3, pp. 419–427 (1962).

[7] Flajolet, P. and Odlyzko, A., The average height of binary trees and other simple trees, *J. Computer & System Sciences*, 25, pp. 171–213 (1982).

[8] Feller, W., *An introduction to probability theory and its applications*, Vol. II., John Wiley & Sons, New York (1971).

[9] Frenk, J., van Houweninge, M. and Rinnooy Kan, A., Order statistics and the linear assignment problem, *Computing*, 39, pp. 165–174 (1987).

[10] Garey, R. and Johnson, D.S., *Computers and intractability: A guide to the theory of NP-completeness*, W.H. Freeman, San Francisco (1979).

[11] Gautschi, W., A computational procedure for incomplete gamma functions, *ACM Trans. on Mathematical Software*, 5, pp. 466–481 (1979).

[12] Goulden, I. and Jackson, D.M., *Combinatorial enumeration*, John Wiley & Sons, New York (1983).

[13] Karp, R., The probabilistic analysis of some combinatorial search algorithms, in: *Algorithms and Complexity*, (ed. J.F. Traub), Academic Press, New York (1976).

[14] Karp, R., Probabilistic analysis of partitioning algorithms for the traveling-salesman problem in the plane, *Math. Oper. Res.*, 2, pp. 209–224 (1977).

[15] Knuth, D., *The art of computer programming, sorting and searching*, Addison-Wesley, Reading (1973).

[16] Lai, T. and Robbins, M., Maximally dependent random variables, *Proc. Nat. Acad. Sci. USA*, 73, pp. 286–288 (1986).

[17] Lai, T. and Robbins, H., A class of dependent random variables and their maxima, *Z. Wahrscheinhich,,* 42, pp. 89–111 (1978).

[18] Loulou, R., Maximal path in random dynamic graphs, *Europ. J. Combinatorics*, 8, pp. 303–311 (1987).

[19] Lueker, G., Optimization problems on graphs with independent random edge weights, *SIAM J. Computing*, 10, pp. 338–351 (1981).

[20] Palmer, E., *Graphical evolution*, John Wiley & Sons, New York (1985).

[21] Papadimitriou, C., Worst-case and probabilistic analysis of a geometric location problem, *SIAM J. Computing*, 10, pp. 542–557 (1981).

[22] Rényi, A., *Probability theory*, North-Holland, Amsterdam (1970).

[23] Szpankowski, W., Some results on asymmetric V-ary tries, *J. Algorithms*, 8, (1988).

[24] Szpankowski, W., On the analysis of the average height of a digital tree: Another approach, Purdue Unviersity CSD TR-646, (submitted to a journal) (1986).

[25] Walkup, D., On the expected value of a random assignment problem, *SIAM J. Computing*, 8, pp. 440–442 (1979).

[26] Weide, B., Random graphs and graph optimization problems, *SIAM J. Computing*, 9, pp. 552–557 (1980).