Algorithmica (1993) 10: 473–500

Algorithmica

Analytic Variations on Quadtrees[1]

Philippe Flajolet,[2] Gaston Gonnet,[3] Claude Puech,[4] and J. M. Robson[5]

Abstract. Quadtrees constitute a hierarchical data structure which permits fast access to multi-dimensional data. This paper presents the analysis of the expected cost of various types of searches in quadtrees—fully specified and partial-match queries. The data model assumes random points with independently drawn coordinate values.

The analysis leads to a class of "full-history" divide-and-conquer recurrences. These recurrences are solved using generating functions, either exactly for dimension $d = 2$, or asymptotically for higher dimensions. The exact solutions involve hypergeometric functions. The general asymptotic solutions rely on the classification of singularities of linear differential equations with analytic coefficients, and on singularity analysis techniques.

These methods are applicable to the asymptotic solution of a wide range of linear recurrences, as may occur in particular in the analysis of multidimensional searching problems.

Key Words. Analysis of algorithms, Multidimensional search, Quadtrees.

Although the worst case is not very impressive,
the quadtree is shown to be efficient
in many practical problems.
S. S. IYENGAR *et al.* [21, p. 73]

1. Introduction.

A classical geometrical search problem consists in finding records (points, elements) that satisfy a suitable condition in a collection of multi-dimensional data (see Samet's book [33] or general references like [3], [17], [21], [28], and [34]). The elements to be retrieved may be specified by several (or all) of their components. If all components are specified in the search, the problem is called a *fully specified* search. Otherwise, we call it a *partial-match* query.

The *quadtree* structure is due to Finkel and Bentley [10]. It can be used to answer both fully specified and partial-match search problems, and it is based on a tree data structure that extends the classical idea of a binary search tree to multidimensional data. The principle, in the case of planar problems,[6] is simply

[1] The work of Philippe Flajolet was supported in part by the Basic Research Action of the E.C. under Contract No. 3075 (Project ALCOM). A preliminary version of this paper has been presented in the form of an extended abstract at the Second Annual Symposium on Discrete Algorithms [13].
[2] Algorithms Project, INRIA, Rocquencourt, F-78153 Le Chesnay, France.
[3] Informatik, E.T.H. Zentrum, CH-8092 Zurich, Switzerland.
[4] LIENS, Ecole Normale Supérieure, 45 rue d'Ulm, F-75005 Paris, France.
[5] Department of Computer Science, Australian National University, Canberra ACT 2601, Australia.
[6] The original quadtree corresponds to records taken from a two-dimensional space, where a subdivision into *quadrants* is used. We employ the term quadtree also for d-dimensional analogs.

Received May 31, 1991; revised October 7, 1991. Communicated by Bernard Chazelle.

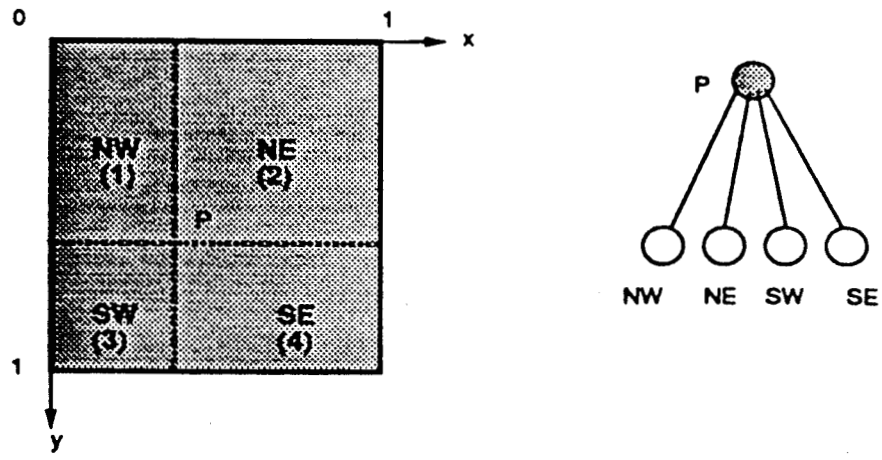

Fig. 1. A point $P = (x, y)$ separates the unit square into four quadrants NW, NE, SW, SE, also numbered 1, 2, 3, 4.

that a point partitions the search space into four quadrants (see Figure 1). When used recursively, this principle leads to a decomposition of the underlying search space into rectangular cells (see Figure 2). A closely related multidimensional tree structure is the k-d tree of Bentley [2].

This paper proposes a thorough analysis of the performances of various types of searches performed on quadtrees built from "random" data. A classical framework of analysis is that of "independent" data, with components of records being independently drawn from some continuous distribution which we may then freely assume to be the uniform distribution over [0, 1].

The quadtree is expected to provide "fast" access properties, and in particular logarithmic cost access to fully specified searches. For instance, in their original

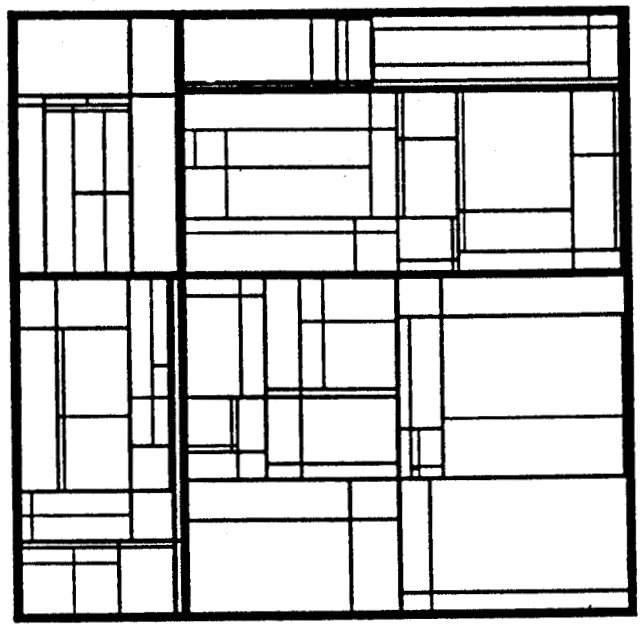

Fig. 2. A quadtree decomposition of the unit square using the principle of Figure 1 recursively, based on 50 points.

paper [10, Table 1], Finkel and Bentley observed by simulations that, for trees of size $n = 1000$ or 10,000, the average cost of a search tends to be about $(0.90 \pm 0.05) \log n$. Gonnet in the first edition in 1984 of the book [17] proposed empirical formulae suggesting $C_n \sim (0.989 \pm 0.004) \log n$ (for dimension $d = 2$) and $C_n \sim (0.662 \pm 0.003) \log n$ (for $d = 3$).

Our asymptotic complexity results are valid for every dimension $d \geq 2$. They are expressed in terms of the number of nodes traversed in a search, more complex measures being amenable to similar analysis techniques. A fully specified search is found to have average cost

$$C_n^{(d)} \sim \frac{2}{d} \log n. \tag{1}$$

(These results are thus in good agreement with the empirical estimates mentioned above.) When comparing the cost of a search in a common (one-dimensional) binary search tree [22] which is $\sim 2 \log n$, we witness a "contraction factor" of $1/d$ for the depth of d-dimensional quadtrees. This represents a sort of global conservation of the search costs (each node in a quadtree contains d fields), a phenomenon independently established in similar contexts by Devroye and Laforest [8], [9] using probabilistic arguments.

One of the main uses of quadtrees is for partial-match queries. In that case, only s out of d coordinates, with $1 \leq s < d$, are specified in a search. First, a simplified model based on the assumption that the quadtree is a perfect tree may be considered. (See [4] and p. 513 of [2] for a similar model of k-d trees.) This leads to considering the recurrence for the cost $\hat{Q}_n^{(s,d)}$ in the perfect tree model

$$\hat{Q}_n^{(s,d)} = 1 + 2^{d-s} \hat{Q}_{n/2^d}^{(s,d)}, \tag{2}$$

since a search in a tree of size n first visits the root and then continues to explore 2^{d-s} trees each of size about $n/2^d$ by the assumption of a perfect tree. The solution of (2) is

$$\hat{Q}_n^{(s,d)} \doteq \Theta(n^{1-s/d}). \tag{3}$$

Said otherwise, a perfect quadtree resembles a perfect grid with meshes of size $n^{-1/d}$.

It turns out that the model (2) provides an unduly optimistic estimate for random data. The exact form of the recurrence for the average search cost $Q_n^{(s,d)}$ is given in Section 2 below. The corrected form of (3) is then found to be

$$Q_n^{(s,d)} = \Theta(n^{1-s/d+\theta(s/d)}), \tag{4}$$

where the correction function $\theta(x)$ in the *exponent* is defined as *the solution* $\theta \in [0, 1]$ *of the equation*

$$(\theta + 3 - x)^x (\theta + 2 - x)^{1-x} - 2 = 0. \tag{5}$$

For instance, when $d = 2$, a partial-match query with one component specified out of two has expected cost

$$Q_n = O(n^{(\sqrt{17}-3)/2}) \approx O(n^{0.56155})$$

as opposed to $O(\sqrt{n})$ which is suggested by the approximate model. This situation resembles the case of k-d trees which has been treated earlier by Flajolet and Puech [15], though the multiplicative constants are naturally different.

The analysis problems that we discuss here start with what may be called *stochastic divide-and-conquer* recurrences. These recurrences on average costs are direct reflections of the recursive search procedures. A typical instance is the recurrence corresponding to path length in a standard quadtree,

$$f_n = n + \sum_{k=0}^{n-1} \xi_{n,k} f_k, \tag{6}$$

where the $\xi_{n,k}$ are related to "splitting probabilities" (see below Lemma 3):

$$\xi_{n,k} = \frac{4}{n}[H_n - H_k] \qquad \text{with} \quad H_n = 1 + \tfrac{1}{2} + \tfrac{1}{3} + \cdots + \frac{1}{n}. \tag{7}$$

The natural approach to recurrences of the form (6) is of course to introduce generating functions. We thus set

$$f(z) := \sum_{n \geq 0} f_n z^n.$$

A recurrence of the form (6), (7) then translates into a linear integral equation, itself equivalent to a *linear differential equation* of order 2. More generally, problems in dimension d lead to differential equations of order d. The analysis of quadtrees follows from their two different routes.

In dimension $d = 2$, the differential equations that we encounter have explicit solutions which invariably invole *hypergeometric functions*, the formulae for partial match being typical. In this way, explicit forms—involving harmonic numbers or binomial coefficients—are available for the complexity analysis of standard quadtrees. Asymptotic forms are derived by elementary or complex asymptotic analysis.

In dimension $d \geq 3$, we no longer find explicit forms of generating functions that would be expressible in terms of known special functions. Our approach is inspired by the corresponding analysis of k-d trees in [15]. The principles on which the analysis is based are:

(i) The nature and location of singularities of a function determine the growth of its coefficients (see, e.g., [14]).

(ii) Singularities of the solution to a linear differential equation

$$\sum_{j=0}^{d} \lambda_j(z) \frac{d^j}{dz^j} f(z) = a(z),$$

arise from singularities of the coefficients $\lambda_j(z)$ and the zeros of $\lambda_d(z)$ in a well-quantified way.

The k-d trees lead to differential systems while quadtrees introduce more naturally integrodifferential equations. However, in both cases, the analysis of generating functions' singularities via differential systems constitutes a fairly general methodology which may be used in order to analyze linear recurrences with coefficients that involve multiple summations and rational function coefficients.

Coming back to quadtrees, we establish here that, not too surprisingly, their expected performances are, as far as orders of growths are concerned, rather close to those of k-d trees. This agrees with some other results of Devroye who established that the height of quadtrees is logarithmic, like for the other search-tree varieties; Devroye proved that for a quadtree of size n, the expected height is asymptotic to

$$\frac{c}{d}\log n, \qquad \text{where} \quad c \approx 4.31107 \quad \text{satisfies} \quad ce^{(1-c)/c} = 2.$$

We may also mention that analyses of quadtrees under different uses, like for representing images or as an access method for data bases, have been given by Yahia *et al.* [30], [38], [26] and Régnier [31].

2. Basic Probabilities and Recurrences. The average-case complexity of divide-and-conquer algorithms is normally expressed by recurrences. For instance, the average number of comparisons C_n needed to sort n data items using the *Quicksort* algorithm satisfies the recurrence [22, equation 5.2.2-18, p. 120]

$$C_n = n + 1 + \frac{2}{n}\sum_{k=0}^{n-1} C_k, \tag{8}$$

and a closely related recurrence [22, equation 6.2.2-4, p. 427] provides the average search cost in a binary search tree of size n. Digital searching leads to recurrences of a different shape, see, for instance, equation 6.3–17, p. 499, of [22].

The general scheme which covers the examples above as well as the quadtree costs is

$$f_n = a_n + \sum_{k=0}^{n-1} \xi_{n,k} f_k. \tag{9}$$

Here f_n is the unknown sequence of costs which is to be determined, a_n is a known (and usually simple) number sequence, and the $\zeta_{n,k}$ are of various forms that reflect, in each case, the probabilities that a problem of size n decomposes into similar subproblems of size k. The form (9) is more complex than the standard divide-and-conquer recurrences of which (2) is a particular example, and we may call it a *stochastic divide-and-conquer recurrence* (Table 1).

Table 1. Various types of stochastic divide-and-conquer recurrences.*

Problem	a_n	$\xi_{n,k}$
Quicksort	$n+1$	$\frac{2}{n}$
Binary search	$\frac{2n}{n+1}$	$\frac{1}{n+1}$
Patricia search	1	$\frac{1}{2^n-2}\binom{n}{k}$
Quadtree path length	n	$\frac{4}{n}(H_n - H_k)$
Quadtree partial match	1	$4\frac{n-k}{n(n+1)}$

* The first three recurrences appear in Knuth's book [22] (on pp. 120, 427, and 479, respectively). The quadtree recurrences appear in Lemmas 3 and 4.

In this section we establish the form of recurrences satisfied by the search costs in a standard quadtree of dimension $d = 2$. Let $\mathbf{U} = [0, 1]^2$ denote the unit square. The probabilistic model of use[7] assumes that n elements are drawn *uniformly* and *independently* from $\mathbf{U}$.

PROPOSITION 1. *Let p_{n_1,n_2,n_3,n_4} be the probability that the four root subtrees of a quadtree built on $n = 1 + n_1 + n_2 + n_3 + n_4$ records have sizes n_1, n_2, n_3, n_4. Then*

$$p_{n_1,n_2,n_3,n_4} = \frac{1}{n \cdot n!} \frac{(n_1+n_2)!\,(n_3+n_4)!\,(n_1+n_3)!\,(n_2+n_4)!}{n_1!\,n_2!\,n_3!\,n_4!}.$$

PROOF. Let $(r_1, r_2, \ldots, r_n)$ be a random element of $\mathbf{U}^n$, and set $r_j = (x_j, y_j)$. The probability sought is

$$(10)\quad p_{n_1,n_2,n_3,n_4} = \binom{n-1}{n_1,n_2,n_3,n_4} \cdot \int_0^1 \int_0^1 [(uv)^{n_1}((1-u)v)^{n_2}(u(1-v))^{n_3}((1-u)(1-v))^{n_4}]\,du\,dv.$$

[7] This model is of course equivalent to assuming simply independent drawings from *any* continuous distribution.

Here $du\ dv$ is the probability that $u \le x_1 < u + du$ and $v \le y_1 < v + dv$. The integral gives the probability that the n_1 elements, $r_2, \ldots, r_{n_1+1}$, are in the first subtree, that the next n_2 elements $r_{n_1+2}, \ldots, r_{n_1+n_2+1}$ are in the second subtree, etc. Finally, the multinomial coefficient represents the number of possible "shufflings" of the $n-1$ elements $r_2, \ldots, r_n$ into four groups of cardinalities n_1, n_2, n_3, n_4.

From the classical Eulerian *beta integral*, see Chapter 6 of [1] or Chapter XII of [36], applied to (10),

$$\int_0^1 x^\alpha (1-x)^\beta \, dx = \frac{\alpha!\,\beta!}{(\alpha+\beta+1)!}, \tag{11}$$

we get the stated form of the splitting probabilities.

We now proceed with determining a recurrence satisfied by the expected values of path length in quadtrees.

DEFINITION 2. The *level* $\lambda(v)$ of a node v in a (quad) tree of root ρ is defined as the number of nodes on the branch connecting v to ρ. (The root itself is thus at level 1.) The (internal) *path length* of a tree t is defined as the sum $\sum_v \lambda(v)$, the sum being extended to all *internal* nodes of t.

We let P_n denote the expectation of path length in a quadtree formed from n random records. Thus P_n/n represents the cost of a random successful search in a random quadtree of size n, the search cost being as usual defined by the number of internal nodes traversed.

LEMMA 3. *The expected value of internal path length P_n in a random quadtree of size n satisfies the recurrence*

$$P_0 = 0, \qquad P_n = n + \frac{4}{n} \sum_{k=0}^{n-1} [H_n - H_k] P_k.$$

PROOF. Let $\pi_{n,k}$ be the probability that the first subtree in a quadtree of n nodes has size k. Then, clearly, we have

$$P_n = n + 4 \sum_{k=0}^{n-1} \pi_{n,k} P_k. \tag{12}$$

The $\pi_{n,k}$ are determined by adapting the argument employed for the full splitting probabilities of Proposition 1. We introduce the intermediate quantities $\varpi_{n,k,l}$ representing the probability that the first subtree has size k while the sizes of the first and third subtrees (west!) add up to l. We thus have

$$\varpi_{n,k,l} = \sum_{\substack{n_1+n_2+n_3+n_4=n-1 \\ n_1=k,\, n_1+n_3=l}} p_{n_1,n_2,n_3,n_4} \quad \text{and} \quad \pi_{n,k} = \sum_{l=k}^{n-1} \varpi_{n,k,l}. \tag{13}$$

The value of $\varpi_{n,k,l}$ is

$$\varpi_{n,k,l} = \binom{n-1}{k,\, l-k,\, n-1-l} \cdot \int_0^1 \int_0^1 (uv)^k (u(1-v))^{l-k} (1-u)^{n-1-l}\, du\, dv \tag{14}$$

$$= \frac{1}{n(l+1)}.$$

The multinomial represents the number of ways in which we can select, out of $n-1$ "places": k places for elements going to the northwest quadrant, $l-k$ places for the southwest quadrant, and the $n-1-l$ remaining places for the east half-plane. The integrand represents the probability that, in a sequence of $n-1$ elements: the first k fall northwest of (u, v), the next $l-k$ fall southwest, and the last $n-1-l$ fall east.

Finally, the explicit form of (14) follows from the beta integrals (11) and the value

$$\pi_{n,k} = \frac{1}{n}[H_n - H_k] \tag{15}$$

is derived from $\varpi_{n,k,l}$ by the second equation of (13).

We now turn to partial-match queries. This is a search on a randomly built tree where only one of the coordinates—the x-coordinate value, say—is specified (and this x-value is taken to be uniform over $[0, 1]$, independently of the data items on which the tree is built). Again cost is measured by the number of internal nodes traversed.

LEMMA 4. *Let Q_n be the expected value of the cost of a partial-match query in a random quadtree of size n. Then Q_n satisfies the recurrence*

$$Q_0 = 0, \qquad Q_n = 1 + \frac{4}{n(n+1)} \sum_{k=0}^{n-1} (n-k) Q_k.$$

PROOF. The probability that $|\mathrm{NW}| = k$, $|\mathrm{SW}| = l-k$, $|\mathrm{E}| = \mathrm{n}-1-l$, and a random partial match with coordinate x is such that x lies west of the root is

$$\varpi^*_{n,k,l} = \binom{n-1}{k,\, l-k,\, n-1-l} \cdot \int_0^1 \int_0^1 (uv)^k (u(1-v))^{l-k} (1-u)^{n-1-l} \underline{u}\, du\, dv \tag{16}$$

$$= \frac{1}{n(n+1)}. \tag{17}$$

Observe here the extra factor of u (underlined) in front of $du\, dv$, which takes into account the probability that a random search hits west. Thus, the recurrence for Q_n is

$$Q_n = 1 + 2 \sum_{0 \le k \le l < n} \varpi^*_{n,k,l} (Q_k + Q_{l-k}). \tag{18}$$

The sum represents the contribution to the cost arising from a search that hits west, the factor 2 takes into account the symmetrical contribution from searches that hit east. The statement follows from (18).

Returning to Lemma 3, we observe that the identity $\varpi_{n,k,l} = 1/(n(l+1))$ in (14) also has an intuitive interpretation: If we consider the sequence $(x_1, x_2, \ldots, x_n)$, the rank of x_1 assumes each of the possible values $1, 2, \ldots, n$ with equal likelihood, i.e., with probability $1/n$. Then the number of elements that fall in the first and second subtrees assumes each of the possible values $l = 0, 1, \ldots, n-1$ with probability $1/n$. Once such a value is fixed, amongst these l elements plus the root, each of the ranks with respect to the y-coordinate of r_1 is equally likely, and has probability $1/(l+1)$. Thus $\varpi_{n,k,l} = 1/(n(l+1))$.

Similar "combinatorial interpretations" are available for (17).

3. Standard Quadtrees in Dimension $d = 2$. In this section we carry out the analysis of search costs in standard quadtrees where the dimension is $d = 2$. Recurrences translate into integrodifferential equations. For $d = 2$, the generating functions can be found explicitly. This leads both to exact and to asymptotic forms for the costs of fully specified searches and partial-match queries. As a uniform measure of costs, we take the *number of internal nodes* traversed in a tree search.

In this and the next section we use a few tools from the theory of linear differential equations for which we refer to the books by Henrici [19] or Wasow [35]. A treatment of hypergeometric functions that suffices for our purposes is to be found in [1] and [36].

PROPOSITION 5. *Let* $P(z) = \sum_{n\geq 0} P_n z^n$ *and* $Q(z) = \sum_{n\geq 0} Q_n z^n$ *be the generating functions of* P_n *and* Q_n. *Then* $P(z)$ *satisfies the second-order equation,* $P(0) = 0$,

$$P(z) = \frac{z}{(1-z)^2} + 4\int_0^z \frac{dt}{t(1-t)} \int_0^t P(u)\frac{du}{1-u}. \tag{19}$$

The function $Q(z)$ *satisfies the differential equation*

$$\frac{d^2}{dz^2}(zQ(z)) = \frac{2}{(1-z)^3} + \frac{4}{(1-z)^2} Q(z), \tag{20}$$

together with the initial conditions $Q(0) = 0$, $Q'(0) = 1$.

PROOF. The proof follows by a direct translation from recurrences to generating functions.

By direct computations, we find

$$P(z) = z + 3z^2 + \tfrac{49}{9}z^3 + \tfrac{295}{36}z^4 + \tfrac{503}{45}z^5 + O(z^6),$$

$$Q(z) = z + \tfrac{5}{3}z^2 + \tfrac{20}{9}z^3 + \tfrac{122}{45}z^4 + \tfrac{2129}{675}z^5 + O(z^6).$$

THEOREM 1. *The expected cost of a positive search—measured by the number of internal nodes traversed—in a quadtree of size $n \geq 1$ is*

$$C_n = \frac{P_n}{n} = \left(1 + \frac{1}{3n}\right)H_n - \frac{n+1}{6n}. \tag{21}$$

The expected cost of a negative search under the same complexity measure in a quadtree of size $n \geq 1$ is

$$C'_n = H_n - \frac{1}{6}\frac{n-1}{n+1}. \tag{22}$$

PROOF. The formula for P_n was initially found by trial-and-error[8] from exact rational number forms of P_n for small n. (The occurrence of the harmonic number is not too unexpected!) Once it has been conjectured, it is a simple matter to verify that the generating function of the P_n as given by (21), namely,

$$P(z) = \frac{1}{3}\frac{2z+1}{(1-z)^2}\log\frac{1}{1-z} + \frac{1}{6}\frac{z^2+4z}{(1-z)^2}, \tag{23}$$

satisfies the second-order integral equation (19).

With regard to C'_n, we have the relation

$$(C'_0 + 1) + (C'_1 + 1) + \cdots + (C'_{n-1} + 1) = nC_n = P_n.$$

Such a relation classically expresses that the search for all records in a tree requires traversing the same nodes as when the record was first inserted into the tree. Thus C'_n results from P_n by differencing: $C'_n = P_{n+1} - P_n - 1$.

COROLLARY 6. *Asymptotically, a random search, either successful or unsuccessful, in a quadtree of size n has average cost* $\log n + O(1)$.

The expected search cost was independently determined by Devroye and Laforest using recurrence manipulations [9]. They further determined an exact

[8] Guessing can be eliminated by systematic reduction to hypergeometric form as we do below for other two-dimensional analyses. Also this problem nowadays falls into a category solvable algorithmically and automatically by computer algebra, see the discussion at the end of Section 4.

form for the variance whose asymptotic equivalent is $\frac{1}{2}\log n$, so that the distribution is concentrated around its mean. Recently, Flajolet and Lafforgue [12] have further obtained an explicit form for the corresponding probability generating function, $\varphi_n(t)$,

$$\varphi_n(u^2) = \frac{1}{4u^2-1}\sum_{j=0}^{n}\binom{2u}{j}\binom{2u-1}{j}\binom{2u-2+n-j}{n-j}. \tag{24}$$

This is based on hypergeometric computations, and from there it is found that (like for standard binary search trees [25]) the distribution of costs is asymptotically Gaussian. The search costs in two-dimensional quadtrees may thus be regarded as fully known, and we turn to partial-match queries.

THEOREM 2. *The expected cost Q_n of a partial-match query in a quadtree of size $n \geq 1$ satisfies*

$$1 + Q_n = \sum_{k=0}^{n}\binom{a-1+n-k}{n-k}\binom{\alpha}{k}\binom{\alpha-1}{k}\frac{1}{k+1}, \tag{25}$$

where α is the root located between 1 and 2 of the equation $\alpha(\alpha+1)=4$; thus $\alpha = (\sqrt{17}-1)/2 \approx 1.56155\ 28128\ 08830$.

PROOF. First we convert the differential equation of $Q(z)$, (20), to standard form:

$$z(1-z)^2\frac{d^2}{dz^2}Q(z) + 2(1-z)^2\frac{d}{dz}Q(z) - 4Q(z) = \frac{2}{1-z}.$$

We observe that a particular solution to this equation is $-1/(1-z)$, and therefore, by considering $y(z) = Q(z) + 1/(1-z)$, we find that $y(z)$ satisfies the homogeneous equation

$$z(1-z)^2\frac{d^2}{dz^2}y(z) + 2(1-z)^2\frac{d}{dz}y(z) - 4y(z) = 0. \tag{26}$$

By general theorems, the only possible singularities of a solution to such an equation are the singularities of the coefficients, and the zeros of the leading coefficient. Thus, the only possible candidates are $z=0$, $z=1$, and $z=\infty$. It is known *a priori*, from the origin of the problem, that the function element $Q(z)$ is regular at 0 and has radius of convergence exactly 1 since its coefficients are polynomially bounded.

Guided[9] by the usual principles of singularity analysis, the local behavior of $Q(z)$, or equivalently $y(z)$, around $z=1$ should be determined in order to derive

[9] An alternative way to arrive at the result is to observe that (26) has three regular singular points and therefore relates to the Riemann P-equation [36].

the asymptotic form of the Q_n. To that purpose, we first try to substitute an asymptotic form $y(z) \sim C/(1-z)^\alpha$ inside (26). The main terms on the left-hand side of (26) are "normally" of order $(1-z)^{-\alpha}$, safe for certain exceptional values of α, where cancellation occurs through the coefficients; we expect precisely these cancellation cases to provide solutions to the differential homogeneous equation. (The left-hand side of (26) must be identically 0.) Proceeding in this way suggests that $y(z) \sim C/(1-z)^\alpha$ with α a root of $\alpha(\alpha+1) = 4$.

To make this precise, we set

$$y(z) = \frac{Y(z)}{(1-z)^\alpha}, \tag{27}$$

with α still kept undetermined at the moment. The function $Y(z)$ satisfies a transformed equation, namely,

(28)

$$z(z-1)^2 \frac{d^2}{dz^2} Y(z) - 2(z-1)(z\alpha - z + 1) \frac{d}{dz} Y(z) + (z(\alpha^2 - \alpha) + 2\alpha - 4)Y(z) = 0.$$

From the preceding discussion, we now fix α to be a root of $\alpha(\alpha+1) = 4$, and we select the largest root, namely, $\alpha = (\sqrt{17} - 1)/2$, since it is the candidate for providing the dominant growth of $y(z)$. In doing so, a term of $(z-1)$ factors out and $Y(z)$ satisfies

$$z(1-z) \frac{d^2}{dz^2} Y(z) + (2 - (2-2\alpha)z) \frac{d}{dz} Y(z) + \alpha(1-\alpha)Y(z) = 0. \tag{29}$$

Equation (29) clearly has three (so-called "regular") singular points at 0, 1, and ∞ and we may compare it with the standard hypergeometric equation.

The classical hypergeometric equation [36, p. 283] involves three parameters, a, b, c. It reads

$$z(1-z) \frac{d^2}{dz^2} F(z) + [c - (a+b+1)z] \frac{d}{dz} F(z) - abF(z) = 0. \tag{30}$$

A formal solution of it defines the classical *hypergeometric function*,

$$F[a, b; c; z] = 1 + \frac{a \cdot b}{c} \frac{z}{1!} + \frac{a(a+1) \cdot b(b+1)}{c(c+1)} \frac{z^2}{2!} + \cdots. \tag{31}$$

Now match the hypergeometric equation (30) with (29) satisfied by $Y(z)$. The correspondence is

$$a = -\alpha, \qquad b = 1 - \alpha, \qquad c = 2. \tag{32}$$

If we adjust the initial conditions with the form of the hypergeometric series, we get

$$Q(z) = \frac{F[-\alpha, 1-\alpha; 2; z]}{(1-z)^{\alpha}} - \frac{1}{1-z}, \tag{33}$$

with $\alpha = (\sqrt{17} - 1)/2$, the hypergeometric function F being defined by (31). (Note: the other solution of the hypergeometric equation has a logarithmic singularity at $z = 0$, so it cannot appear in a generating function.)

By the binomial expansion and the hypergeometric expansion, we have

$$\frac{1}{(1-z)^{\alpha}} = \sum_{n=0}^{\infty} \binom{\alpha + n - 1}{n} z^n,$$

$$F[a, b; 2; z] = \sum_{n=0}^{\infty} \binom{-a}{n}\binom{-b}{n} \frac{z^n}{n+1}. \tag{34}$$

This determines an explicit convolution form of the coefficients of $Q(z)$, as described in (33). The statement of the theorem follows.

From the generating function form (33) of $Q(z)$, detailed asymptotic information on the coefficients Q_n is available. By the general principles of *singularity-analysis* techniques [14] that we review now, the asymptotic form of Q_n is determined by the aysmptotic properties of $Q(z)$ at its singularity $z = 1$.

SINGULARITY ANALYSIS. This method is based on two principles. First, if we examine coefficients of standard functions that are singular at $z = 1$, we observe that functions that get larger around $z = 1$ have larger coefficients. Let $[z^n]f(z)$ denote the coefficient of z^n in $f(z)$. Approximating the binomial coefficients, we find

$$[z^n] \frac{1}{(1-z)^{\alpha}} = \frac{n^{\alpha-1}}{\Gamma(\alpha)} + O(n^{\alpha-2}). \tag{35}$$

Next, it can be proved under a variety of conditions that the type of estimate (35) also holds for functions only known *asymptotically* at $z = 1$,

$$[z^n] O\left(\frac{1}{(1-z)^{\beta}}\right) = O(n^{\beta-1}). \tag{36}$$

One set conditions ensuring the validity of the "transfer" of (36) is that the expansion of the function holds in an extended domain of the complex plane.

The combination of (35) and (36) shows that once a singular expansion of a function has been obtained, the asymptotic form of its Taylor coefficients is known. Thus, under the analytic continuation conditions of [14], we have the implication

$$f(z) = \frac{C}{(1-z)^\alpha} + O\left(\frac{1}{(1-z)^\beta}\right) \quad \Rightarrow \quad [z^n]f(z) = \frac{C}{\Gamma(\alpha)} n^{\alpha-1} + O(n^{\alpha-2} + n^{\beta-1}).$$

THEOREM 3. *The expected cost of a partial-match query in a quadtree of size $n \geq 1$ satisfies asymptotically*

$$Q_n \sim \gamma n^{\alpha-1}, \qquad \text{where} \quad \gamma = \frac{1}{2}\frac{\Gamma(2\alpha)}{\Gamma(\alpha)^3}, \tag{37}$$

with $\alpha = (\sqrt{17} - 1)/2$. Numerically $\gamma \approx 1.59509\ 90958\ 29715$.

PROOF. First, by a classical identity of Gauss, we have

$$F[a, b; c; 1] = \frac{\Gamma(c)\Gamma(c-a-b)}{\Gamma(c-a)\Gamma(c-b)}, \tag{38}$$

whenever the real part $\Re(c - a - b) > 0$, and $c \neq 0, -1, -2, \ldots$ (see Section 14.11 of [36] or Article 15.1.20 of [1]). Thus, we find from (33) that

$$Q(z) \sim \frac{\gamma^*}{(1-z)^\alpha} \qquad (z \to 1), \tag{39}$$

with

$$\gamma^* = F[-\alpha, 1-\alpha; 2; 1] = \frac{\Gamma(2\alpha+1)}{\Gamma(2+\alpha)\Gamma(1+\alpha)}.$$

That asymptotic expansion is easily found to hold true in an extended domain of the complex plane since the hypergeometric function only has algebraic or logarithmic branch points. Thus, by singularity analysis [14], we are able to "transfer" the asymptotic relation on $Q(z)$ into a corresponding asymptotic form of Q_n, namely,

$$Q_n \sim \gamma^* \frac{n^{\alpha-1}}{\Gamma(\alpha)}.$$

The statement of the theorem thus follows with $\gamma = \gamma^*/\Gamma(\alpha)$.

A refinement of this argument leads to a full asymptotic expansion for the Q_n.

COROLLARY 7. *Define the asymptotic series in n,*

$$\varphi(\theta, n) \sim 1 + \sum_{k=1}^{\infty} \frac{(\theta-1)^3 \cdots (\theta-k)^3}{(2\theta)\cdots(2\theta-k+1)} \frac{\theta}{\theta-k} \cdot \frac{(-1)^k}{k!} \cdot \frac{1}{(n+\theta-1)\cdots(n+\theta-k)}.$$

Then

$$1 + Q_n \sim \frac{1}{2}\frac{\Gamma(2\alpha)}{\Gamma(\alpha)^2}\binom{n+\alpha-1}{n}\varphi(\alpha, n) + \frac{1}{2}\frac{\Gamma(2\bar{\alpha})}{\Gamma(\bar{\alpha})^2}\binom{n+\bar{\alpha}-1}{n}\varphi(\bar{\alpha}, n), \tag{40}$$

with $\alpha = (-1 + \sqrt{17})/2$, *and* $\bar{\alpha}$ *the conjugate of* α, $\bar{\alpha} = (-1 - \sqrt{17})/2$.

PROOF. By using relations between hypergeometric functions at z and at $1 - z$, a complete singular expansion of $Q(z)$ results. From the known expansions, see Section 14.53 of [36] or Article 15.3.6 of [1], we find

$$Q(z) + (1-z)^{-1} = \gamma^*(1-z)^{-\alpha}F[-\alpha, 1-\alpha; -2\alpha; 1-z] + \gamma^{**}(1-z)^{-\bar{\alpha}}F[-\bar{\alpha}, 1-\bar{\alpha}; -2\bar{\alpha}; 1-z],$$

where γ^{**} derives from γ^* by changing α to $\bar{\alpha}$. That convergent expansion can in turn be transferred termwise to coefficients.

The form (40) provides an asymptotic expansion (that is divergent!) of Q_n. The asymptotic scale involves inverses of descending "factorials" of $n + \alpha$ and $n + \bar{\alpha}$.

We have thus found a new expansion of Q_n as a sum of two purely divergent formal ${}_3F_2$-hypergeometric forms. The quality of the approximation that we obtain by retaining the first four terms of the expansion (40)—these terms all come from $\varphi(\alpha, n)$—is already quite exceptional; for $n = 1, 10, 100, 1000$, the absolute error is of order respectively $10^{-2}, 10^{-6}, 10^{-9}, 10^{-12}$. The error is tiny, even for $n = 1$, while the series is divergent!

4. Higher Dimensions. In this section we examine the cost of various searches in quadtrees for data taken in higher-dimensional spaces. The recurrences involve more complicated splitting probabilities and the generating function equations have integral forms that reduce to linear differential equations of order d, when the dimension is equal to d. The results are less explicit than in the case $d = 2$, but orders of growth can still be precisely quantified although, in the case of partial match, the multiplicative constants no longer appear to have closed forms (to the best of our knowledge!).

We use in an essential manner singularity-analysis techniques. We are thus led to analyzing generating function solutions to ODE locally around their dominant singularity at $z = 1$.

- The case of a fully specified search illustrates a situation in which the dominant asymptotic behavior at $z = 1$ comes from the inhomogeneous term in the differential equation.
- The case of a partial-match query corresponds to a situation where the dominant asymptotic contribution comes from solutions to the associated homogeneous equation.

In both cases we use a modest amount of the theory of singular points of ordinary linear differential equations as may be found in books by Henrici [19] or Wasow [35].

SINGULAR DIFFERENTIAL SYSTEMS. By a classical theorem, the singularities of a homogeneous *linear* differential equation or system can only arise from singularities of the coefficients. For systems, a particularly important case occurs when the coefficient matrix is meromorphic and the singularity under consideration is only a *simple* pole. The singularity is then called *regular*. If the singularity is normalized to occur at $z = 1$, a fundamental result implies that solutions of the form

$$\frac{1}{(1-z)^{\alpha}} \cdot \sum_{k=0}^{\infty} c_k (1-z)^k$$

exist. By substituting inside the original equations, we need to obtain complete cancellation. It is then seen that only a finite number of possibilities exist for α; these are solutions of a polynomial equation which is known as the *indicial equation*. The process could be called a method of "indeterminate exponents," and further full expansions follow by the usual technique of indeterminate coefficients. Finally, inhomogeneous equations are treated by means of quadratures once solutions to the associated homogeneous equations have been determined (the process is known under a nickname, the "variation-of-constant" method).

LEMMA 8. *Let P_n denote the expected internal path length in a d-dimensional quadtree of size n. The P_n satisfy the recurrence*

$$(41) \quad P_n = n + 2^d \sum_{k=0}^{n-1} \pi_{n,k} P_k \qquad \text{with} \quad \pi_{n,k} = \frac{1}{n} \sum_{\mathscr{L}} \frac{1}{(l_1+1)(l_2+1)\cdots(l_{d-1}+1)},$$

where the summation is over all sequences $(l_1, l_2, \ldots, l_d)$ with the condition $\mathscr{L}$ being $n > l_1 \geq l_2 \geq \cdots \geq l_{d-1} \geq l_d = k$.

The generating function $P(z) = \sum_{n \geq 0} P_n z^n$ *satisfies the integral equation*

$$P(z) = \frac{z}{(1-z)^2} + 2^d \mathbf{J}^{d-1} \mathbf{I} P(z), \tag{42}$$

where the operators **I**, **J** *are defined by*

$$\mathbf{I}f(z) = \int_0^z f(t) \frac{dt}{1-t}, \qquad \mathbf{J}f(z) = \int_0^z f(t) \frac{dt}{t(1-t)}.$$

PROOF. The quantity $\pi_{n,k}$ represents the probability that the first root subtree has size k. The term $(n(l_1 + 1) \cdots (l_{d-1} + 1))^{-1}$ gives the probability that the root splits the file into $\langle l_1, n - l_1 - 1 \rangle$ elements according to the first dimension, and the l_1 elements are themselves split into $\langle l_2, l_1 - l_2 \rangle$ according to the next dimension, etc. The complete justification by means of integral calculations is entirely similar to that of Lemma 3. The translation to generating functions from there is immediate.

We make a digression at this stage regarding splitting probabilities. The probability $\pi_{n,k}$ that a tree of size n has its first root-subtree of size k (see (41)) admits alternative forms. One of them involves generalized harmonic numbers and it extends the case $d = 2$ of (15); for $d = 3$, for instance, we find

$$\pi_{n,k} = \frac{1}{2n} ([H_n - H_k]^2 + [H_n^{(2)} - H_k^{(2)}]), \qquad \text{where} \quad H_n^{(d)} = \sum_{j=1}^{n} \frac{1}{j^d}.$$

A concise form valid for all $d \geq 2$ is

$$\pi_{n,k} = \binom{n-1}{k} \sum_{j=0}^{n-1-k} \binom{n-1-k}{j} \frac{(-1)^j}{(k+j+1)^d}.$$

This form is also discussed by Laforest in her thesis together with related results of interest [24]. The equivalence between the various forms relies on elementary properties of generalized harmonic numbers. (A somewhat related problem of geometric probabilities is discussed in [7].) Binomial expressions (in the style of Proposition 1) of full splitting probabilities are available in all cases.

THEOREM 4. *The expected cost* $C_n^{(d)}$ *of a fully specified search in a d-dimensional quadtree of size n satisfies*

$$C_n^{(d)} = \frac{2}{d} \log n + \lambda_d + O\left(\frac{\log n}{n} + n^{2\cos(2\pi/d) - 2}\right), \tag{43}$$

for some real constant λ_d.

PROOF. The rough idea of the proof is that (42) behaves as a perturbation of a simpler equation that can be solved explicitly. This fact relies on the observation that the two functionals $\mathbf{I}f(z)$ and $\mathbf{J}f(z)$ act as "singularity transformers" (around

the singularity $z = 1$) in a similar way, as far as main orders of growth are concerned. The proof proceeds in three steps.

(A) Consider first the *simplified homogeneous equation* in which **J** is replaced by **I**:

$$y(z) - 2^d \mathbf{I}^d y(z) = 0. \tag{44}$$

If we try a solution of the form $(1 - z)^{-\alpha}$, we find the condition

$$\alpha^d - 2^d = 0, \tag{45}$$

which is the indicial equation associated with (44). Equation (44) is in fact an Euler equation; it has exact solutions of the form ($j = 0, \ldots, d - 1$)

$$y_j(z) = (1 - z)^{-2\omega^j} \qquad \text{with} \quad \omega = e^{2i\pi/d}. \tag{46}$$

(B) The *simplified inhomogeneous equation associated* with (44) which "approximates" the original equation (42) is

$$g(z) - 2^d \mathbf{I}^d g(z) = \frac{z}{(1 - z)^2}. \tag{47}$$

Standard resolution techniques are best expressed in terms of systems rather than equations, a viewpoint that we now adopt.

Equation (47) can be split using intermediate variables, and, e.g., for $d = 3$, we have an equivalent system

$$g_1 - 2\mathbf{I}g_2 = \frac{z}{(1 - z)^2}, \qquad g_2 - 2\mathbf{I}g_3 = 0, \qquad g_3 - 2\mathbf{I}g_1 = 0,$$

with $g(z) \equiv g_1(z)$. By differentiation, we are thus led to the vectorial system

$$\frac{d}{dz}\mathbf{g} = \frac{1}{1 - z}\mathbf{A}\mathbf{g} + \mathbf{w}, \tag{48}$$

with $\mathbf{g} = (g_1, \ldots, g_d)^{\mathrm{T}}$ the unknown vector, and $g(z) \equiv g_1(z)$. The matrix **A** involved in the singular part is such that $\frac{1}{2}\mathbf{A}$ is a circular permutation matrix,

$$\tfrac{1}{2}(\mathbf{A})_{i,j} = \delta_{j,\,1+i \bmod d},$$

with $\delta_{u,v}$ the Kronecker symbol. Finally,

$$\mathbf{w} = (w_1, 0, \ldots, 0)^{\mathrm{T}} \qquad \text{with} \quad w_1(z) = \frac{(1 + z)}{(1 - z)^3} = \frac{d}{dz}\frac{z}{(1 - z)^2}. \tag{49}$$

A fundamental matrix $\mathbf{W}$ for the system is by definition a nonsingular $d \times d$ matrix that satisfies the homogeneous matrix system

$$\frac{d}{dz}\mathbf{W} = \frac{1}{1-z}\mathbf{A}\cdot\mathbf{W}.$$

One such matrix is obtained directly by adapting the solutions (46) of the scalar Euler equation:

$$(\mathbf{W})_{i,j} = \frac{\omega^{(i-1)(j-1)}}{(1-z)^{2\omega^{j-1}}}. \tag{50}$$

In other word, this fundamental matrix decomposes into the product of a (Discrete Fourier Transform!) Vandermonde matrix and a diagonal matrix of singular parts:

$$\mathbf{W} = \Omega\cdot\Delta, \qquad (\Omega)_{i,j} = \omega^{(i-1)(j-1)}, \quad \text{and} \quad \Delta = \operatorname{diag}((1-z)^{-2\omega^{j-1}})_{j=1\ldots n}.$$

The inhomogeneous system (48) is then solved by the matrix form of the variation-of-constant method [19, p. 99], now that all solutions to the homogeneous equation are known. The classical formula reads

$$\mathbf{g}(z) = \mathbf{W}(z)\cdot\mathbf{W}^{-1}(0)\cdot\mathbf{g}(0) + \mathbf{W}(z)\cdot\int_0^z \mathbf{W}^{-1}(t)\cdot\mathbf{w}(t)\,dt. \tag{51}$$

The matrix $\mathbf{W}$ is nonsingular since the Vandermonde matrix Ω is itself nonsingular. Computing the inverse $\mathbf{W}^{-1}$ reduces to the corresponding problem for the special matrix Ω, and we find

$$(\mathbf{W}^{-1})_{i,j} = \frac{1}{d}\,\omega^{-(i-1)(j-1)}(1-z)^{2\omega^{i-1}}. \tag{52}$$

A little computation using the forms (52) and (50), the cost function $\mathbf{w}(z)$ in (49), and the initial value $\mathbf{g}(0) \equiv 0$ inside the variation-of-constant formula (51) shows that the solution to (47) satisfies

$$g(z) = \frac{2}{d}\frac{1}{(1-z)^2}\log\frac{1}{1-z} + \frac{C'}{(1-z)^2} + R(z),$$

where

$$R(z) = O((1-z)^{-2\omega}) \quad \text{as} \quad z \to 1 \tag{53}$$

for some explicitly computable constant C'. The logarithm occurs in the integral of (51) because of "resonances" between the homogeneous solution $(1-z)^{-2}$ (or its counterpart $(1-z)^2$ inside $\mathbf{W}^{-1}$) and the inhomogeneous term $O((1-z)^{-2})$; the factor $2/d$ then arises as the product of the coefficient $1/d$ present in $\mathbf{W}^{-1}$ and the coefficient 2 of the inhomogeneous term,

$$w_1(z) = \frac{2}{(1-z)^3} - \frac{1}{(1-z)^2} \sim \frac{2}{(1-z)^3}.$$

Thus we are able to determine ultimately the leading coefficient $2/d$ because the treatment of (48) could be made explicit enough.

(C) We finally return to the *exact equation* satisfied by $P(z)$:

$$f(z) - 2^d \mathbf{J}^{d-1} \mathbf{I} f(z) = \frac{z}{(1-z)^2}. \tag{54}$$

Let $g(z)$ be the solution to the approximate inhomogeneous equation (47) which is described by (53). The induced equation for $h(z) = f(z) - g(z)$ is such that its homogeneous part is that of the original equation while we attain a reduction in order of growth for the inhomogeneous term. More precisely, substituting $f = g + h$ inside (54) and subtracting the equation satisfied by g, we find

$$h(z) - 2^d \mathbf{J}^{d-1} \mathbf{I} h(z) = 2^d [\mathbf{J}^{d-1} - \mathbf{I}^{d-1}] \mathbf{I} g(z). \tag{55}$$

The right-hand side $U(z)$ is found to satisfy $U(z) = O((1-z)^{-1} \log(1-z))$ as $z \to 1$ by repeated integration.

The homogeneous equation corresponding to (55) has the same indicial equation as the approximate equation (44), namely, $\alpha^d - 2^d = 0$. We are thus in a situation similar to our previous paragraph, except that the inhomogeneous term is now small.

Thus, the solution to (55) satisfied by the difference $h(z)$ is a sum of a general homogeneous solution which is by necessity of the form

$$C''(1-z)^{-2} + O((1-z)^{-2\omega})$$

and of a particular inhomogeneous solution, given by the variation-of-constant method, which is found to be of the same asymptotic order as $U(z)$, so that

$$h(z) = C''(1-z)^{-2} + O((1-z)^{-2\omega}) + O((1-z)^{-1} \log(1-z)). \tag{56}$$

Altogether, we have thus found from (53) and (56) that

$$P(z) = g(z) = \frac{2}{d} \frac{1}{(1-z)^2} \log \frac{1}{1-z} + \frac{C}{(1-z)^2} + O((1-z)^{-2\omega}) + O((1-z)^{-1} \log(1-z)) \tag{57}$$

for some constant C. The singular expansion of $P(z)$ has been completed. From there, the behavior of P_n, hence that of $C_n^{(d)} = P_n/n$, follows by direct singularity analysis [14].

We note that the main term of $(2/d) \log n$ in Theorem 4 was determined independently by Devroye and Laforest [8], [9] using probabilistic methods; they also showed that the distribution of costs is concentrated around the mean. Using techniques of [16], Flajolet and Lafforgue have shown that the distribution of costs is actually Gaussian in the limit.

At this point, it is worth summarizing the asymptotic process that we have employed here. Given a differential equation (or some equivalent integral form) of order d,

$$\Psi f(z) = w(z), \tag{58}$$

with $f(z)$ the unknown function, we first determine the allowable growths for the solutions to the homogeneous equation $\Psi f = 0$, by substituting the form $(1 - z)^{-\alpha}$; this leads to a polynomial equation for α, the indicial equation.

In the generic[10] case (no repeated roots, no roots differing by an integer), a collection of d different solutions to the homogeneous equation of the form $(j = 1, \ldots, d)$

$$Y_j(z) = (1 - z)^{-\alpha_j} A_j(z), \tag{59}$$

with A_j analytic at 1, is obtained.

A particular solution to the inhomogeneous solution is obtained by the variation-of-constant method. It is composed of the sum of a finite number of elements of the form

$$B(z)(1 - z)^{-\alpha} \int_0^z C(t)(1 - t)^{\beta} w(t)\, dt \tag{60}$$

for some analytic functions B, C and numbers α, β that satisfy the indicial equation.

Complete solutions can then be composed from homogeneous solutions (59) and inhomogeneous forms (60). The dominant asymptotic behavior near a singularity can then be found by inspection. For instance, in the path-length analysis, the indicial equation is $\alpha^d - 2^d = 0$; the dominant exponent is $\alpha = 2$, which corresponds to the fact that all homogeneous solutions grow at most like $(1 - z)^{-2}$. Given some inhomogeneous term $w(z)$ corresponding to a basic cost function in the ODE, the variation-of-constant method provides a solution with dominant asymptotic growth of the form

$$\frac{1}{d}(1 - z)^{-2} \int_0^z (1 - t)^2 w(t)\, dt. \tag{61}$$

Concerning path length, it is the case that everything is driven by the inhomogeneous term.

---

[10] In other cases, integral powers of logarithmic terms appear that complicate the picture without altering it in an essential way.

This general discussion serves us for treating partial-match queries (where homogeneous terms are found to predominate).

LEMMA 9. *Let Q_n represent the expected number of node traversals in a partial-match query of a random d-dimensional quadtree containing n points. Then $Q_0 = 0$ and, for $n \geq 1$, the Q_n satisfy the recurrence*

$$Q_n = 1 + 2^d \sum_{k=0}^{\infty} \pi^*_{n,k} Q_k \tag{62}$$

with

$$\pi^*_{n,k} = \frac{1}{n(n+1)} \sum_{\mathscr{L}} \left[\frac{1}{(l_1+2)\cdots(l_{s-1}+2)}\right] \cdot \left[\frac{1}{(l_{s+1}+1)\cdots(l_{d-1}+1)}\right],$$

where the summation takes place over all sequences $(l_1, l_2, \ldots, l_d)$ satisfying the condition $\mathscr{L}$ above.

The generating function $Q(z) = \sum_{n\geq 0} Q_n z^n$ satisfies the integral equation

$$z \frac{d^2}{dz^2}(zQ(z)) = \frac{2z}{(1-z)^3} + 2^d \frac{1}{1-z} \mathbf{J}^{s-1} \frac{z}{1-z} \mathbf{I}^{d-s-1} Q(z). \tag{63}$$

PROOF. The lemma provides the basic recurrences that hold true for higher-dimensional partial-match search. Its proof is a direct adaptation of the methods employed for the computation of geometric probabilities in Section 2, see especially Lemma 4.

THEOREM 5. *The expected cost $Q_n^{(s,d)}$ of a partial-match query in a d-dimensional quadtree of size n, with s coordinates specified, satisfies asymptotically*

$$Q_n^{(s,d)} \sim \gamma_{s,d} n^{\alpha - 1} \tag{64}$$

for some constant $\gamma_{s,d} \neq 0$ and α the root between 1 and 2 of the equation

$$\alpha^{d-s}(\alpha+1)^s = 2^d.$$

In other words, we have

$$Q_n^{(s,d)} \sim \gamma_{s,d} n^{1 - s/d + \theta(s/d)},$$

where the function $\theta(x)$ is defined as the solution $0 < \theta < 0.07$ of the equation

$$(\theta + 3 - x)^x (\theta + 2 - x)^{1-x} - 2 = 0. \tag{65}$$

PROOF. We only briefly sketch it since we described the general asymptotic process in great length earlier, and since the analytic behavior of our generating functions is akin to what we find for k-d trees (for example, the indicial equation is identical).

The linear integral equation (63) is equivalent to a differential equation of order d as seen by repeatedly taking derivatives:

(66)
$$\mathbf{I}^{s+1-d}\frac{1-z}{z}\mathbf{J}^{1-s}z(1-z)\frac{d^2}{dz^2}(zQ(z)) - 2^dQ(z) = \mathbf{I}^{s+1-d}\frac{1-z}{z}\mathbf{J}^{1-s}\frac{2z}{(1-z)^2},$$

where the *differential* operators $\mathbf{I}^{-1}$ and $\mathbf{J}^{-1}$ satisfy

$$\mathbf{I}^{-1}f(z) = (1-z)\frac{d}{dz}f(z), \qquad \mathbf{J}^{-1}f(z) = z(1-z)\frac{d}{dz}f(z).$$

If we try to satisfy approximately the integral equation (63) or its equivalent differential form (66), by a function $1/(1-z)^\alpha$, we find the indicial equation

(67)
$$\alpha^{d-s}(1+\alpha)^s = 2^d.$$

By referring to the vectorial form of the equation that we obtain upon setting

$$g_j(z) = (1-z)^{j-1}\frac{d^{j-1}}{dz^{j-1}}Q(z),$$

we check that (66) has a regular singularity at $z = 1$.

Thus, the homogeneous equation (66) has linearly independent solutions of the form

(68)
$$\frac{1}{(1-z)^{\alpha_j}}\sum_{k=0}^{\infty} c_k^{(j)}(1-z)^k,$$

for $j = 1, \ldots, d$, where the α_j are roots of the indicial equation (67). A complete set of d solutions is obtained in this way provided that there are no repeated roots—this was established for all d in [15]—and that no roots differ by an integer—which was checked using computer algebra for all $d \leq 10$. Our discussion now assumes the latter property to be satisfied. (If not, the proof can be salvaged at the expense of some minor complications with logarithmic terms, see again [15].)

Amongst the functions (68), it is the one with $\Re(\alpha_j)$ maximal that gives the dominant contribution around $z = 1$. This corresponds to the unique solution α of the indical equation (67) that belongs to the real interval $(1, 2)$.

A discussion that we omit (see Theorem 4)) based on the variation-of-constant method shows that, contrary to the case of a fully specified search, the inhomogeneous terms introduce contributions that are asymptotically negligible at $z = 1$, being of order $O((1-z)^{-1})$. (The discussion can also be based on special solutions, see additional remarks below.)

We thus find for some constant C the estimate

$$Q(z) \sim \frac{C}{(1-z)^{\alpha}}, \qquad \alpha^{d-s}(1+\alpha)^s = 2^d, \quad \alpha \in (1,2). \tag{69}$$

That asymptotic form is then transferred to Q_n by the usual methods of singularity analysis, which provides

$$Q_n \sim \frac{C}{\Gamma(\alpha)} n^{\alpha-1}.$$

The auxiliary fact that the coefficient C is nonzero may be established by the same argument as for k-d trees [15, Lemma 5]. Finally, the bounds on the correction function $\theta(x)$ result from elementary real analysis.

We conclude this section with some remarks on special solutions to our differential equations.

First, the homogeneous differential equation (corresponding to the integral forms (19), (42)) satisfied by the path-length generating function P has a special solution, which for an arbitrary dimension d is

$$P^*(z) = \frac{1 + (2^d - 2)z}{(1-z)^2}. \tag{70}$$

This function $P^*(z)$ is the generating function for the number of leaves in quadtrees; for instance, when $d = 2$, its Taylor coefficient of rank n is $1 + 3n$, which represents the number of leaves in any quaternary tree with n internal nodes. There is a reason for this particular solution: from its tree definition, $P^*(z)$ satisfies the modified integral equation

$$P^*(z) - 2^d \mathbf{J}^{d-1} \mathbf{I} P^*(z) = 1,$$

hence also the associated homogeneous differential equation. This fact then "explains" the explicit elementary solution for path length (and similar parameters) in dimension 2, and in general it entails a reduction of order, so that the path-length generating function satisfies, furthermore, an equation of order $d - 1$ over $\mathbb{C}(x)$.

Manuel Bronstein, in a private communication [6], has found similar rational solutions for the inhomogeneous equations satisfied by the Q-functions. Using dedicated algorithms to compute rational solutions to linear ODEs (see [5]) and

the Maple system for symbolic manipulations, Bronstein found the particular solutions

$$Q^*_{s,d}(z) = \frac{1}{2^{d-s}-1}\frac{1}{z-1} \qquad \text{when} \quad s < d-1,$$

(71)

$$Q^*_{s,d}(z) = \frac{1}{2^{d-2}}\frac{1+(2^{d-2}-1)z}{z-1} \qquad \text{when} \quad s = d-1.$$

The combinatorial significance of these solutions does not, however, appear obvious to us at present.

5. Conclusions. Multidimensional search problems may lead to intricate stochastic divide-and-conquer recurrences. As originally seen with k-d trees and further demonstrated here, a powerful method consists in studying these recurrences via generating functions.

Quite clearly, any suitably "additive" parameter of quadtrees in dimension ≥ 2 can be analyzed by the methods described here. In order to support this claim we offer a final analysis, the search for records with smallest y-coordinate in a quadtree (lowest points). Concerning k-d trees, the analysis was given in Guibas' problem in the *Journal of Algorithms* [18] for dimension $d = 2$, and it is discussed in Puech's thesis [29] for all $d \geq 2$. We obtain here:

THEOREM 6. *The expected cost R_n of finding the lowest point in a quadtree with n internal nodes satisfies:*

For $d = 2$, exactly,

$$1 + R_n = \sum_{k=0}^{n} \binom{\alpha}{k}\binom{\alpha-1}{k}\binom{\alpha-1+n-k}{n-k} \qquad \textit{with} \quad \alpha = \sqrt{2}.$$

For $d = 2$, asymptotically,

$$R_n \sim \frac{\Gamma(2\alpha)}{\alpha\Gamma(\alpha)^3}\, n^{\alpha-1} \qquad \textit{with} \quad \alpha = \sqrt{2}.$$

For $d > 2$, asymptotically,

$$R_n \sim \gamma_d n^{\alpha-1} \qquad \textit{with} \quad \alpha = 2^{1/d}.$$

PROOF. (Sketch!) For dimension 2, a search always descends in the two south subtrees. The recurrence is thus very similar to the one for partial match, see (18), except that the symmetry factor 2 in front of the sum is absent. This leads to an

equation for the generating function of costs which, for arbitrary dimension d, reads

$$R(z) = \frac{z}{1-z} + 2^{d-1}\mathbf{J}^{d-1}\mathbf{I}R(z).$$

The indicial equation is then

$$\alpha^d - 2^{d-1} = 0.$$

The hypergeometric solution for $d = 2$ is found to be

$$R(z) + \frac{1}{1-z} = \frac{F[-\sqrt{2}, 1-\sqrt{2}; 1; z]}{(1-z)^{\sqrt{2}}},$$

from which the rest follows. In higher dimensions, the discussion is entirely similar to that of partial match.

The techniques of this paper have been applied recently by Hoshi and Flajolet [20] to determine the storage occupation of paged quadtrees used as indexes for two-dimensional data. As a special case, the proportion of leaves in a randomly grown standard quadtree is asymptotically $(4\pi^2 - 39)$. These results have been in turn complemented by analyses of Labelle and Laforest [23] who employed a parallel recurrence approach in order to obtain the asymptotic proportion of nodes in a quadtree having one child as $(24\zeta(3) - 156\zeta(2) + 228)$. Finally, as mentioned earlier, see (24) and Section 4, the distribution of levels of nodes in quadtrees can be studied in a similar vein (by a method of "regular perturbation") and proved to be Gaussian in the limit [12], which refines the concentration result of Devroye and Laforest [9].

We now have available a method of considerable generality that can be used to study a large number of linear recurrences[11] with variable coefficients, both in homogeneous and inhomogeneous cases.

It applies to many recurrences of the full history type with coefficients that are rational functions of n and k, which leads to linear differential equations with analytic coefficients. These ODEs can be analyzed *locally* in the neighborhood of their singularities, using the classical results from the theory of linear differential equations. The singular expansions so obtained are then to be "transferred" back to the original number sequence, either by means of the method of singularity analysis in the case of regular singular points or by means of saddle-point methods in the case of irregular singularities (see, e.g., [11] for an irregular singularity and an explicit generating function).

[11] The proper class entertains close ties with the class of *holonomic* functions introduced by Zeilberger in a powerful approach [39].

In both the regular and the irregular cases, everything rests on the fact that singularities have been classified [19], [35] and are formed of elements of one of two types;

$$(72)\qquad \mathbf{R}\colon\ (1-z)^{-\alpha}\log^{k}\frac{1}{1-z}\qquad\text{and}\qquad \mathbf{I}\colon\ (1-z)^{-\alpha}\exp((1-z)^{-p/q}).$$

Methods of contour integration apply well here in this systematic context of isolated singularities (see, e.g., [14], [27], and references therein, or [32] for a survey). The induced asymptotic forms for the coefficients are of two corresponding types which are schematically

$$(73)\qquad \mathbf{R}'\colon\ n^{\alpha-1}\log^{k} n\qquad\text{and}\qquad \mathbf{I}'\colon\ n^{\alpha'}\exp(n^{p'/q'}).$$

The general method suggested by this discussion then constitutes an alternative to the direct treatment of recurrences by the theory of difference equations that was developed by G. D. Birkhoff, and is examined extensively by Wimp and Zeilberger in [37].

Acknowledgments. The authors are grateful to Michèle Loday-Richaud and Manuel Bronstein for several useful suggestions. Thanks to a referee for a very careful scrutiny of the paper, and to another one for encouraging remarks.

References

[1] Abramowitz, M., and Stegun, I. A. *Handbook of Mathematical Functions.* Dover, New York, 1973. A reprint of the tenth National Bureau of Standards edition, 1964.

[2] Bentley, J. L. Multidimensional binary search trees used for associative searching. *Communications of the ACM*, **18**(9) (Sept. 1975), 509–517.

[3] Bentley, J. L., and Friedman, J. H. Data structures for range searching. *ACM Computing Surveys*, **11**(4) (1979), 397–409.

[4] Bentley, J. L., and Stanat, D. F. Analysis of range searching in quad trees. *Information Processing Letters*, **3**(6) (July 1975), 170–173.

[5] Bronstein, M. On solutions of linear ordinary differential equations in their coefficient field. Technical Report 152, Department Informatik, ETH, January 1991.

[6] Bronstein, M. Rational solution to differential equations related to quadtrees. Private communication, February 1991.

[7] Buchta, C. On the average number of maxima in a set of vectors. *Information Processing Letters*, **33** (Nov. 1989), 63–65.

[8] Devroye, L. Branching processes in the analysis of the heights of trees. *Acta Informatica*, **24** (1987), 277–298.

[9] Devroye, L., and Laforest, L. An analysis of random d-dimensional quad trees. *SIAM Journal on Computing*, **19** (1990), 821–832.

[10] Finkel, R. A., and Bentley, J. L. Quad trees, a data structure for retrieval on composite keys. *Acta Informatica*, **4** (1974), 1–9.

[11] Fishburn, P. C., Odlyzko, A. M., and Roberts, F. S. Two-sided generalized Fibonacci sequences. Preprint, 1989.

[12] Flajolet, P., and Lafforgue, T. Search costs in quadtrees and singularity perturbation asymptotics. 1993, submitted.

[13] Flajolet, P., Gonnet, G., Puech, C., and Robson, J. M. The analysis of multidimensional searching in quad-trees. *Proceedings of the Second Annual ACM–SIAM Symposium on Discrete Algorithms*, Philadelphia, 1991, SIAM, Philadelphia, PA, pp. 100–109.

[14] Flajolet, P., and Odlyzko, A. M. Singularity analysis of generating functions. *SIAM Journal on Discrete Mathematics*, **3**(2) (1990), 216–240.

[15] Flajolet, P., and Puech, C. Partial match retrieval of multidimensional data. *Journal of the ACM*, **33**(2) (1986), 371–407.

[16] Flajolet, P., and Soria, M. Gaussian limiting distributions for the number of components in combinatorial structures. *Journal of Combinatorial Theory*, Series A, **53** (1990), 165–182.

[17] Gonnet, G. H., and Baeza-Yates, R. *Handbook of Algorithms and Data Structures: in Pascal and C*, second edn. Addison-Wesley, Reading, MA, 1991.

[18] Guibas, L. Problems. *Journal of Algorithms*, **3**(4) (1982), 362–380. ⟨Problem 80-6⟩, from the Stanford 1979 Algorithms Qualifying Examination. Solution by Eric S. Rosenthal, pp. 368–371.

[19] Henrici, P. *Applied and Computational Complex Analysis.* Wiley, New York, 1977. Three volumes.

[20] Hoshi, M., and Flajolet, P. Page usage in a quadtree index. *BIT*, **32** (1992), 384–402.

[21] Iyengar, S. S., Rao, N. S. V., Kashyap, R. L., and Vaishnavi, V. K. Multidimensional data structures: review and outlook. *Advances in Computers*, **27** (1988), 69–119.

[22] Knuth, D. E. *The Art of Computer Programming*, vol. 3: Sorting and Searching. Addison-Wesley, Reading, MA, 1973.

[23] Labelle, G., and Laforest, L. Variations combinatoires autour des arborescences hyperquaternaires. Technical Report, LACIM, UQAM, Montreal, Nov. 1991.

[24] Laforest, L. Étude des arbres hyperquaternaires. Technical Report 3, LACIM, UQAM, Montreal, Nov. 1990. (Author's Ph.D. thesis at McGill University.)

[25] Louchard, G. Exact and asymptotic distributions in digital and binary search trees. *RAIRO Theoretical Informatics and Applications*, **21**(4) (1987), 479–495.

[26] Mathieu, C., Puech, C., and Yahia, H. Average efficiency of data structures for binary image processing. *Information Processing Letters*, **26** (Oct. 1987), 89–93.

[27] Odlyzko, A. M., and Richmond, L. B. Asymptotic expansions for the coefficients of analytic generating functions. *Aequationes Mathematicae*, **28** (1985), 50–63.

[28] Preparata, F. P., and Shamos, M. I. *Computational Geometry, An Introduction.* Springer-Verlag, New York, 1985.

[29] Puech, C. *Méthodes d'analyse de structures de données dynamiques.* Doctorat ès sciences, Université de Paris Sud, Orsay, 1984.

[30] Puech, C., and Yahia, H. Quadtrees, octrees, hyperoctrees: a unified approach to tree data structures used in graphics, geometric modeling and image processing. *Proceedings of the First ACM Symposium on Computational Geometry*, Baltimore, 1985, pp. 272–280.

[31] Régnier, M. Analysis of grid file algorithms. *BIT*, **25** (1985), 335–357.

[32] Salvy, B. Asymptotique automatique et fonctions génératrices. Ph.D. thesis, École Polytechnique, 1991.

[33] Samet, H. *The Design and Analysis of Spatial Data Structures.* Addison-Wesley, Reading, MA, 1990.

[34] Sedgewick, R. *Algorithms*, second edn. Addison-Wesley, Reading, MA, 1988.

[35] Wasow, W. *Asymptotic Expansions for Ordinary Differential Equations.* Dover, New York, 1987. A reprint of the Wiley edition, 1965.

[36] Whittaker, E. T., and Watson, G. N. *A Course of Modern Analysis*, fourth edn. Cambridge University Press, Cambridge, 1927. Reprinted 1973.

[37] Wimp, J., and Zeilberger, D. Resurrecting the asymptotics of linear recurrences. *Journal of Mathematical Analysis and Applications*, **111** (1985), 162–176.

[38] Yahia, H. Analyse des structures de données arborescentes représentant des images. Doctorat de troisième cycle, Université de Paris Sud, Orsay, Dec. 1986.

[39] Zeilberger, D. A holonomic approach to special functions identities. *Journal of Computational and Applied Mathematics*, **32** (1990), 321–368.