The Annals of Applied Probability
1997, Vol. 7, No. 2, 523–544

DETERMINING THE MAJORITY: THE BIASED CASE

BY PHILIPPE CHASSAING

Université Henri Poincaré

We are given a set of n elements, some of them red, the others blue, but their colors are hidden. We are to determine the composition of this set, or to determine an element of the majority color, by making pairwise comparisons of elements from which we obtain the information "the colors of these two elements are the same," or "they are different." Let τ_n, respectively, μ_n, be the optimal average number of comparisons needed to solve these two problems. We give an explicit expression of the limit of τ_n/n, respectively, of μ_n/n, in terms of the probabilities of being red or blue. We also discuss quasi-optimal algorithms in both cases: when these probabilities are known and when they are unknown.

1. Introduction. Given a set of n elements $\{x_1, x_2, x_3, \ldots, x_n\}$, some of the elements being red, the others blue, we consider successively two problems: we want to exhibit an element of the majority color, and we want to determine the composition of the set. For this aim, we can make pairwise comparisons: are the colors of the elements x_k and x_m equal or different? Notice that we are somewhat colorblind. We are unable to determine the majority color, or the color of a given element, but our problem is different: we have to give the answer "I do not know the color of x_k, but I know that x_k belongs to the majority color" for at least one among the elements of the set, and in the composition problem, we have to obtain the final conclusion that there are k elements of some color and $n-k$ of the other color. The composition problem turns out to be tightly related to the majority one, in the sense that its solution gives the solution of the majority problem (see Section 9 for explanations).

The motivation for the majority problem comes from system diagnosis. According to Schmeichel, Hakimi, Otsuka and Sullivan (1990), "in a set U of n units (processors, modules, etc.) at most t are faulty, and an external observer wishes to identify the faulty units. The observer acquires information by requesting the results of certain tests performed by one unit upon another; e.g. $u_i \in U$ might be asked to determine if u_j is faulty or not. If u_i is fault-free then the test performed by u_i is assumed reliable; if u_i is faulty however, u_i may find u_j faulty or fault-free, regardless of the actual condition of u_j." An algorithm that finds the faulty units exists if and only if $t < n/2$ [see Preparata, Metze and Chien (1967)]: we need to be sure that some unit is fault-free to rely on its diagnosis of the remaining units, but we cannot be

Received June 1995; revised October 1996.
AMS 1991 *subject classifications*. Primary 68Q25, 90C15; secondary 93E20, 90C40.
Key words and phrases. Graph, connected component, martingale, quasi-optimal algorithm, Bellman principle.

sure of that, since a faulty unit can behave exactly as a fault-free one. However there are configurations of test results in which the assumption that some particular unit u_0 is faulty would entail that a majority of units would be faulty, too: if we know that $t < n/2$, we are then sure that u_0 is fault-free, and we can rely on its diagnosis.

The best algorithm [in the worst case, see Hakimi and Schmeichel (1984)] already known for system diagnosis is quite similar to the worst case optimal algorithm given by Alonso, Reingold and Schott (1993) for the majority problem described in the first paragraph, and also similar to the average case quasi-optimal algorithm given by Alonso, Reingold and Schott (1994) for the majority problem, in the case where the probability p of being red is equal to the probability q of being blue. However, the assumption of equality between p and q is not consistent with the assumption $t < n/2$ and with the general belief that faulty units are not a majority. The algorithm proposed by Hakimi and Schmeichel (1984) could thus perform poorly on the average in the case $p \neq q$. Our first motivation is that an average case quasi-optimal algorithm for the majority problem, in the case $p \neq q$, could lead us to an algorithm for system diagnosis performing better on the average than the algorithm proposed by Hakimi and Schmeichel.

Another motivation is that the average optimality is held to be more significant than the worst case optimality, at least for algorithms one uses very often, due to the law of large numbers, but the average optimality is also usually held to be much more difficult to establish [see Knuth (1973), page 217]. Actually, there are a few fundamental problems in which an average quasi-optimal algorithm is known: the sorting problem [see Knuth (1973)], the selection problem [see Cunto and Munro (1984)] or the majority problem in the case $p = q$ are some examples. A tight lower bound is found, using an information theory argument for the first problem and using combinatorics for the two other problems. Here a tight lower bound is provided by quite different tools, that is, by martingale arguments.

We assume that each element has probability p (respectively, q) of being red (respectively, blue), independently of the others. In this paper, T will denote the number of pairwise comparisons required to obtain the composition of the set, or, if we study the majority problem, T will denote the number of comparisons required to produce an element of the majority color (in this last case, when n is even and when there is no majority, T will denote the number of comparisons required to be sure that there is no majority). We look for an average optimal algorithm, that is, an algorithm minimizing the expected number of comparisons $E[T]$.

2. The graph and its connected components. Of course, one can keep track of comparisons on a graph whose vertices would be the n elements of our set, by adding, at each comparison, a marked edge between the two elements just compared (see Figure 1): G_t will denote the graph obtained after the comparison number t. If, at time t, we compare some element x_i to

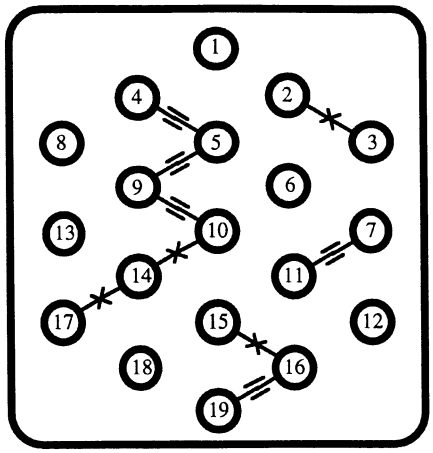

FIG. 1.

some x_j, the track of this comparison is kept by

$$G_t = G_{t-1} \cup \{(i, j)\}.$$

Obviously at each step we add an edge joining two distinct connected components of the previous graph. Otherwise we would ask a question whose answer could be computed with the help of preceding answers (for instance, in Figure 1, we do not need to compare x_4 and x_{17} to see that their colors are the same). Incidentally, any algorithm that satisfies the previous rule requires a number T of comparisons that it is less than or equal to $n - 1$.

An algorithm is described by a map associating with a graph G_t the next edge to be drawn, according to this algorithm. Let τ_n (respectively, μ_n) denote the optimal average cost for the composition problem (respectively, the majority problem).

3. Notation and results. In this paper $\lfloor x \rfloor$ will denote the lower integer part of the real number x. We assume, without loss of generality, that $p \geq q$, and we set

$$\rho = q/p,$$

so that $0 \leq \rho \leq 1$.

In Section 4 we point out some invariances of the problem. In Section 5 we give the equation satisfied by μ_n (respectively, τ_n). Unfortunately, it does not furnish a closed form expression of μ_n (respectively, τ_n), nor of the optimal algorithm. In the following sections, we prove the following asymptotics of μ_n and τ_n.

THEOREM 1.

$$n\phi(\rho) - 1 \leq \tau_n \leq n\phi(\rho) + \left\lfloor \frac{\log n}{\log 2} \right\rfloor,$$

where

$$\phi(\rho) = \frac{1-\rho}{2(1+\rho)}\left(\frac{1+\rho}{1-\rho} + \frac{1+\rho^2}{2(1-\rho^2)} + \frac{1+\rho^4}{4(1-\rho^4)} + \frac{1+\rho^8}{8(1-\rho^8)} + \cdots\right).$$

THEOREM 2. *For any* $\alpha > 1/2$,

$$\mu_n = \frac{1+\rho}{2}\phi(\rho)n + o(n^\alpha).$$

Note that ϕ is well defined at 1, as is $(1-\rho)/(1-\rho^n)$, and that $\phi(1) = 2/3$, respectively $\phi(0) = 1$. From tedious computations, it turns out that ϕ is decreasing, while $((1+\rho)/2)\phi$ is increasing, as expected. Quasi-optimal algorithms corresponding to Theorems 1 and 2 appear in Sections 7 and 9, respectively. The residual terms are by no means optimal when $\rho < 1$: for Theorem 1, it is proven in Section 7, Corollary 2, that $\lfloor \log n/\log 2 \rfloor$ can be replaced by a constant $C(\rho)$. For Theorem 2, my conjecture is that the residual term can be improved to $O(1)$, by making the quasi-optimal algorithm more adaptive (see Section 11, Concluding remarks). However, in the uniform case ($p = q$), previous work shows that the hypothesis $\alpha > 1/2$ is optimal.

The first result on the subject, as far as I know, is due to Saks and Werman (1991); they proved that, in the worst case, at least $n - \nu(n)$ comparisons were necessary, in which $\nu(n)$ is the number of 1-bits in the binary representation of n. A simpler proof was given by Alonso, Reingold and Schott (1993). Alonso, Reingold and Schott (1994) proved that in the uniform case, the average complexity for the majority problem is at least

$$\frac{2}{3}n - \sqrt{\frac{8n}{9\pi}} - O(1).$$

They also describe an algorithm achieving an average complexity of

$$\frac{2}{3}n - \sqrt{\frac{8n}{9\pi}} + O(\log n).$$

The question of the biased case seems natural, since, in system diagnosis, the probability of being defective is implicitly assumed to be less than $1/2$. The beautiful proofs of Alonso, Reingold and Schott are pure combinatorics, and I had to introduce quite different tools, coming from the optimal control of discrete stochastic processes, to solve the biased case.

4. Some invariances of the problem. The graph G_t being given, how do we choose the next comparison or the next edge? Of course, the pair of vertices does not matter, but only the pair of connected components joined by the edge. Which connected components do we compare? Because of the symmetric role played by the n elements, it does not matter if we choose one connected component or another, provided that they have the same composition.

For any connected component c of G_t, let us denote by $K(c)$ the number of elements (of c) that belong to the majority color of c and $k(c)$ the number belonging to the minority color. The composition of c is described by the pair $(K(c), k(c))$. Let the *value* of the connected component c be denoted by $v(c)$ and defined by

$$v(c) = K(c) - k(c).$$

For instance, in Figure 1, the connected component $\{4, 5, 9, 10, 14, 17\}$ has composition $(5, 1)$ and value 4.

Actually, the optimal choice of the connected components to be compared depends on their compositions only through their values: if we replace a connected component with composition $(b + k, b)$ and value k with a connected component with composition $(k, 0)$—that is, if we put aside b elements of each color—common sense suggests that it makes no difference for the majority problem. A proof of this fact is not required to establish our main results, but we give some ingredients of this proof that are essential in later sections. In the first place, we have the following proposition.

PROPOSITION 1. *If the t-th comparison involves two connected components, c and c', with compositions $(a + k, a)$ and $(b + m, b)$, respectively, where k and m are positive, the result is a connected component whose value V satisfies*

$$P(V = k + m) = \frac{1 + \rho^{k+m}}{(1 + \rho^k)(1 + \rho^m)}$$

and

$$P(V = |k - m|) = \frac{\rho^k + \rho^m}{(1 + \rho^k)(1 + \rho^m)}.$$

Since these probabilities do not depend on the compositions, but only on the values k and m, of c and c', respectively, they give also the distribution of V under the condition that the values of c and c' are respectively k and m.

PROOF. The compositions $(a + k, a)$ (respectively, $(b + m, b)$) of c (respectively, c') can be read on the graph G_{t-1}; they are known when we decide the next comparison. The probability that the majority colors of the two components are the same, conditioned on the compositions of c and c', is the quotient of the probability that $a + b + k + m$ elements of $c \cup c'$ exactly are blue (respectively, red), and that among these blue (respectively, red) elements, exactly $a + k$ are taken from c, that is,

$$\binom{2a + k}{a}\binom{2b + m}{b}(p^{a+k+b+m}q^{a+b} + q^{a+k+b+m}p^{a+b}),$$

by the probability of observing the compositions $(a + k, a)$ and $(b + m, b)$, that is,

$$\binom{2a + k}{a}\binom{2b + m}{b}(p^{a+k}q^a + q^{a+k}p^a)(p^{b+m}q^b + q^{b+m}p^b).$$

This quotient gives

$$\frac{1+\rho^{k+m}}{(1+\rho^k)(1+\rho^m)}.$$

The second case follows at once. □

The previous proposition indicates that the ratio of probabilities of being red or blue, for the majority of a connected component, is ρ^k when the component's value is k, while it is of course ρ when the connected component has a single element—and also when its value is 1. Actually, we have the proposition:

PROPOSITION 2. *If some connected component, say c, has a positive value k, its majority color is red with probability*

$$\frac{1}{1+\rho^k},$$

and it is blue with probability

$$\frac{\rho^k}{1+\rho^k}.$$

The easy proof is omitted. The fact that the values are some kind of sufficient statistic appears also in the stopping condition for the majority problem: let $(c_i)_{1\le i\le n-t}$ denote the $n-t$ connected components of G_t after step t, with composition $(K_i,k_i)_{1\le i\le n-t}$, and associated values $(v_i)_{1\le i\le n-t}$. We are sure that the majority color of the connected component c_i is the same as in the whole set if and only if

$$K_i+\sum_{j\neq i}k_j > k_i+\sum_{j\neq i}K_j$$

since in the worst case, the majority color in c_i is the minority color in the other connected components. This condition depends on the values, as it can be written

$$\nu_i > \sum_{j\neq i}\nu_j. \tag{4.1}$$

In fact an element of the majority color is known if at least one connected component satisfies the condition above, that is, if

$$\max \nu_j > \sum_{j=1}^{n-t}\nu_j - \max \nu_j.$$

Thus the stopping condition depends only on the values of the connected components. The additive cost is 1 at each step, until the stopping condition is satisfied, and thus does not depend on the compositions of the connected components either. We have gathered all the ingredients of a formal proof of the fact that there exists an optimal decision rule depending only on the values.

For the composition problem, only the stopping condition is different: we stop when the values of all the connected components are 0, except for one value which would be, let us say, k. The composition is then

$$\left(\frac{n+k}{2}, \frac{n-k}{2}\right).$$

Note that it never helps (for either problem) to use a comparison involving a component of value 0.

5. The optimality equation and the optimal algorithm. Let $C_k(t)$ be the number of connected components of G_t with value k, and let

$$C_t = (C_k(t))_{k \geq 1}.$$

To a sequence $u = (u_n)_{n \geq 1}$ of nonnegative integers, we associate

$$d(u) = \sup\{k | u_k > 0\},$$

$$|u| = \sum_{k=1}^{+\infty} k u_k \quad \text{and} \quad \#u = \sum_{k=1}^{+\infty} u_k.$$

Set

$$E = \{u | d(u) < +\infty\}.$$

We denote by e_k the element of E whose kth term is equal to 1, the other terms being equal to 0, and by e_0 the null element. The state of the system is exhaustively described by G_t but, according to Section 4, the useful information is contained in C_t. We shall say that the state of the system is $u \in E$ if

$$C_t = u.$$

We say that we choose decision (i, j), if we decide to compare a connected component with value i to a connected component with value j. The set of possible decisions in state u, say $A(u)$, is the set of couples of positive integers (i, j) such that $u - e_i - e_j$ still belongs to E. As a consequence of decision (i, j), the next state will be

$$T^+_{i,j}(u) = u - e_i - e_j + e_{i+j}$$

with probability

$$p_+(i, j) = \frac{1 + \rho^{i+j}}{(1 + \rho^i)(1 + \rho^j)},$$

according to Proposition 1, and it will be

$$T^-_{i,j}(u) = u - e_i - e_j + e_{|i-j|},$$

with probability

$$p_-(i, j) = \frac{\rho^i + \rho^j}{(1 + \rho^i)(1 + \rho^j)}. \tag{5.1}$$

For the majority problem, the terminating set is

$$\Delta = \{u | d(u) > |u|/2\} \cup \{e_0\}.$$

The terminal state e_0 occurs when all the connected components have value 0, or equivalently when we are sure that there is no majority. Let $\mu(u)$ be the optimal average number of comparisons for the majority problem, starting from state u: μ_n is also $\mu(ne_1)$. The optimality equation of stochastic dynamic programming [see Bertsekas (1987)] can then be written as follows.

PROPOSITION 3. *When u does not belong to Δ,*

$$\mu(u) := 1 + \min\Big\{p_+(i,j)\mu\big(T^+_{i,j}(u)\big) + p_-(i,j)\mu\big(T^-_{i,j}(u)\big)\Big|(i,j) \in A(u)\Big\}.$$

Of course, μ is zero on Δ. The optimal decision in state u is, as usual, any couple of the set

$$\arg\min\Big\{p_+(i,j)\mu\big(T^+_{i,j}(u)\big) + p_-(i,j)\mu\big(T^-_{i,j}(u)\big)\Big|(i,j) \in A(u)\Big\}.$$

The system of Proposition 3, though nonlinear, is easy to solve recursively, being "triangular": for an appropriate total ordering $\lhd$ of E, we have

$$T^+_{i,j}(u) \lhd u, \qquad T^-_{i,j}(u) \lhd u,$$

and $\{v \mid v \lhd u\}$ is finite. We set:

$$u \lhd v$$

iff

$$|u| < |v| \qquad \text{or} \qquad \big\{|u| = |v| \text{ and } u_K > v_K \text{ holds true for } K = \max\{k|u_k \neq v_k\}\big\}.$$

For instance, when $\rho = 1$, starting with the smaller elements u of E ($|u|$ even and u not in Δ), Proposition 3 entails that

$$\mu(2e_1) = \mu(2e_n) = 1, \qquad \mu(2e_1 + e_2) = 3/2, \qquad \mu(4e_1) = 9/4,$$
$$\mu(e_1 + e_2 + e_3) = 3/2, \qquad \mu(3e_1 + e_3) = 7/4, \qquad \mu(3e_2) = \mu(3e_n) = 1,$$

and, finally, the first case where not all decisions are optimal is:

$$\mu(2e_1 + 2e_2) = 3/2,$$

in which the (unique) optimal decision is $(2, 2)$.

Note that the set $\{u \in E \mid |u| = n\}$ is the set of partitions of the integer n [see Andrews (1976)]. According to the work of Hardy and Ramanujan (1918), the computational effort needed to solve the Bellman equation for $\mu(ne_1)$ is prohibitive, since

$$\log\big[\#\{v|v \lhd ne_1\}\big] \approx \sqrt{n}\,.$$

We have not been able to guess a closed form expression of μ. For the composition problem, the optimal cost $\tau(u)$ is similarly given by the following.

PROPOSITION 4. *We have $\tau(u) = 0$ when u belongs to the stopping set $\Delta' = \{u \mid \#u \leq 1\}$, and*

$$\tau(u) := 1 + \min\Big\{p_+(i,j)\tau\big(T^+_{i,j}(u)\big) + p_-(i,j)\tau\big(T^-_{i,j}(u)\big)\Big|(i,j) \in A(u)\Big\}$$

when u does not belong to Δ'.

6. A lower bound for τ_n. Assume that f is a bounded nonnegative function on $[0,1]$, satisfying the following two properties:

$$\forall (x,y) \in [0,1]^2, \text{ s.t. } x \vee y \neq 0,$$

$$\text{(i)} \qquad 1 + \frac{1+xy}{(1+x)(1+y)} f(xy) + \frac{x+y}{(1+x)(1+y)} f\left(\frac{x \wedge y}{x \vee y}\right) \geq f(x) + f(y),$$

$$\text{(ii)} \qquad \forall\, x \in [0,1], \qquad 1 + \frac{1+x^2}{(1+x)^2} f(x^2) \geq 2f(x).$$

Such a function is necessarily bounded by 1. Examples are positive constants less than or equal to $2/3$, and also $(1-x)/(1+x)$.

PROPOSITION 5. *For any function f satisfying* (i) *and* (ii), *the following relation holds true*:

$$\tau_n \geq f(\rho) n - 1.$$

PROOF. In this proof, we assume that we are given some rule R mapping G_t to the new edge to be drawn. Let M_t be defined by

$$M_t = (T \wedge t) + \sum_{k=1}^{+\infty} \lambda_k C_k (T \wedge t).$$

in which

$$\lambda_k := f(\rho^k).$$

Let F_{t-1} be the σ field generated by the G_s, $s < t$. If t is less or equal than T and if, according to R, the t-th comparison A_t is to be between two connected components with two different values, let us say $A_t = (k,m)$, we have

$$E(M_t - M_{t-1} | F_{t-1}, A_t) = 1 + \frac{1+\rho^{k+m}}{(1+\rho^k)(1+\rho^m)} \lambda_{k+m} + \frac{\rho^k + \rho^m}{(1+\rho^k)(1+\rho^m)} \lambda_{|k-m|} - \lambda_k - \lambda_m,$$

but if they have the same size k, we have

$$E(M_t - M_{t-1} | F_{t-1}, A_t) = 1 + \frac{1+\rho^{2k}}{(1+\rho^k)^2} \lambda_{2k} - 2\lambda_k. \tag{6.1}$$

Then we deduce the lemma from (i) and (ii).

LEMMA 1. *The process M_t is a submartingale.*

As a consequence,

$$E[M_t] \geq E[M_0].$$

For $t = n$, it follows that

$$E[T] \geq \lambda_1 n - \sum_{k=1}^{+\infty} \lambda_k E\big[C_k(T)\big].$$

From the definition of λ_k and the fact that all the $C_k(T)$ but at most one are zero, and that one (if any) equals 1, we obtain

$$E(T) \geq f(\rho)n - E\big[f(\rho^{|C_T|})\big], \tag{6.2}$$

for any algorithm R. This ends the proof of Proposition 5. □

The first inequality of Theorem 1 follows from

PROPOSITION 6. *The function* ϕ *satisfies* (i) *and* (ii).

PROOF. Actually, ϕ satisfies equality in (2):

$$2\phi(x) = 1 + \frac{1 + x^2}{(1 + x)^2}\phi(x^2). \tag{6.3}$$

In order to prove that ϕ satisfies (1), let us make the change of variable $x = e^{-2\alpha}$. We have then

$$\psi(\alpha) = \phi(e^{-2\alpha}) = \sum_{n \geq 0} 2^{-n-1}\psi_{2^n}(\alpha),$$

in which

$$\psi_n(\alpha) := \tanh \alpha \coth n\alpha.$$

A sufficient condition for (i) to hold true is that, for any nonnegative numbers x and y and any positive integer n,

$$1 + \frac{1 + \tanh x \tanh y}{2}\psi_n(x + y) + \frac{1 - \tanh x \tanh y}{2}\psi_n(x - y) \geq \psi_n(x) + \psi_n(y). \tag{6.4}$$

According to the suggestion of L. Alonso, we write (6.4) in the following form:

$$1 + \frac{\tanh x}{2}\Theta(nx, ny) + \frac{\tanh y}{2}\Theta(ny, nx) \geq 0,$$

in which

$$\Theta(x, y) = \coth(x + y) + \coth(x - y) - 2\coth x.$$

From the convexity of coth, we obtain the positivity of $\Theta(x, y)$ only when x is larger than y. A little more work gives

$$\Theta(x, y) = -\frac{2\coth x}{\sinh^2 x(\coth^2 x - \coth^2 y)},$$

from which we can write (6.4) in the form

$$\frac{\chi_n(x) - \chi_n(y)}{\coth^2 nx - \coth^2 ny} \geq 0.$$

with

$$\chi_n(x) = \frac{1 - \psi_n(x)}{\sinh^2 nx}.$$

But it turns out that both χ_n and $\coth^2$ are decreasing. □

This is the first part of the proof of Theorem 1.

7. A quasi-optimal algorithm for the composition problem. The only bounded solution to the functional equation (6.3) turns out to be ϕ. The bound given by ϕ would be sharp for an algorithm comparing only components with equal values, if such an algorithm would exist: we see from (6.1) that M_t would then be a martingale, and $E[T]$ would reach the lower bound given by (6.2). We shall now describe an algorithm in which comparisons involving components with different values are very scarce. We derive then the second inequality of Theorem 1.

For any integer x, let $g(x), h(x)$ be defined by

$$g(x) = \left\lfloor \frac{x}{2} \right\rfloor, \qquad h(x) = x - 2g(x) = 1_{x \text{ is odd}}.$$

The quasi-optimal algorithm has I main steps, I being a random variable satisfying

$$I \leq 1 + \left\lfloor \frac{\log n}{\log 2} \right\rfloor. \tag{7.1}$$

In the first step, we have Y_1 $(= n)$ connected components with value 1 and we do $g(Y_1)$ pairwise comparisons, obtaining $g(n)$ components, Y_2 among them having value 2, and the $g(n) - Y_2$ others having value 0, plus eventually a connected component with value 1, if n is odd. That is, the first step produces Y_2 components with value 2, and $h(Y_1)$ components with value 1, plus some useless components with value 0 that we forget, at a cost of $g(Y_1)$ pairwise comparisons. For instance, in Figure 2–4 $Y_1 = 19$, $g(Y_1) = 9$, $h(Y_1) = 1$, $Y_2 = 5$.

In the second step, if $g(Y_2)$ is not 0, we perform $g(Y_2)$ pairwise comparisons between the components with value 2, and at the end of this step, we have produced Y_3 components with value 4, $g(Y_2) - Y_3$ components with value 0—that we forget—and $h(Y_2)$ components with value 2, and so on. We see in Figure 5 that $g(Y_2) = 2$, $h(Y_2) = 1$ and $Y_3 = 1$.

The number I of steps is the first index k such that $g(Y_k) = 0$, and (7.1) is a consequence of the fact that $Y_{k+1} \leq Y_k/2$. After the Ith step, we have $h(Y_k)$ components with value $2^{k-1}(1 \leq k \leq I)$, Z of the $h(Y_k)$ being equal to 1, the others 0, and a set of components with value 0 (useless for the composition problem). If $Z = 0$, we conclude that there is no majority. If $Z \neq 0$, doing $Z - 1$ comparisons between the Z nonzero components, we obtain a unique connected component with value $S \neq 0$, and the same set of useless components with value 0 as we had before. We conclude that the majority leads by

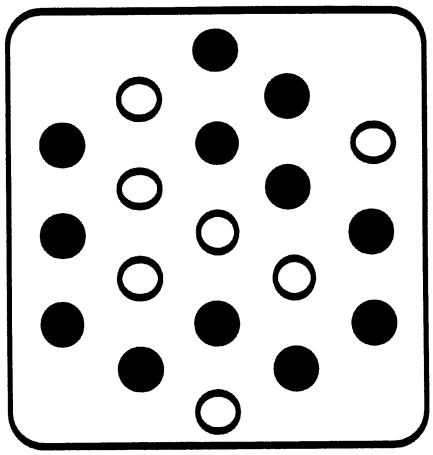

FIG. 2.

S units so that the composition is $((n + S)/2, (n - S)/2)$. In Figures 5 and 6, we see that $I = 3$, $h(Y_3) = h(Y_2) = h(Y_1) = 1$ and thus $Z = I$. In Figure 7 the two last comparisons, (6, 19) and (6, 18), give $S = 5$, and a composition (12, 7), as expected, at a total cost of 13 comparisons.

Set Y_k to be zero for $k > I$. The total number of comparisons is thus

$$T = Z - 1 + \sum_{k \geq 1} g(Y_k).$$

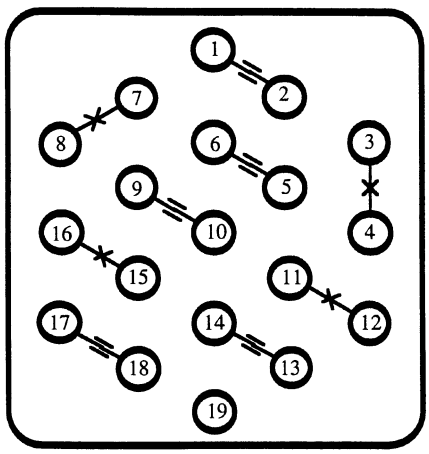

FIG. 3.

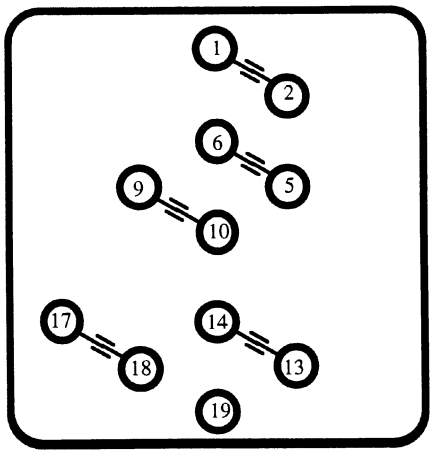

FIG. 4.

We have

$$Z = \sum_{k \geq 1} h(Y_k),$$

$$Z \leq I \leq 1 + \left\lfloor \frac{\log n}{\log 2} \right\rfloor.$$

For $k \geq 2$, the conditional law of Y_k given Y_{k-1} is the binomial law with parameters $g(Y_{k-1})$ and $p_+(2^{k-2}, 2^{k-2})$. Thus, using $g(x) \leq x/2$, we obtain

$$E[Y_k] \leq \tfrac{1}{2} p_+(2^{k-2}, 2^{k-2}) E[Y_{k-1}]$$

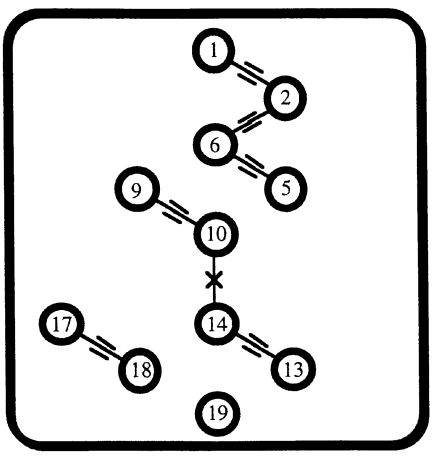

FIG. 5.

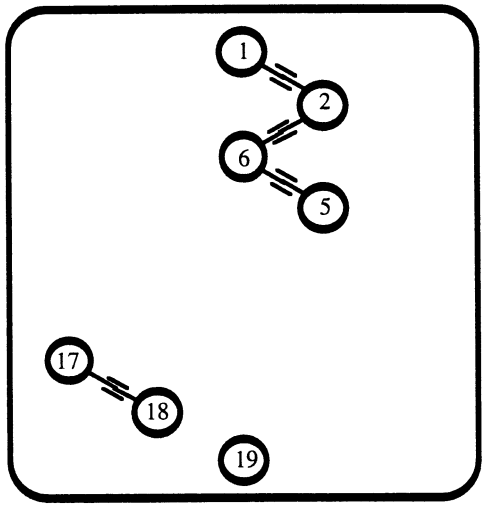

FIG. 6.

and

$$E[g(Y_k)] \le \frac{n}{2^k} \prod_{j=2}^{k} p_+(2^{j-2}, 2^{j-2})$$

$$= \frac{n}{2^k} \frac{1-\rho}{1+\rho} \frac{1+\rho^{2^{k-1}}}{1-\rho^{2^{k-1}}}.$$

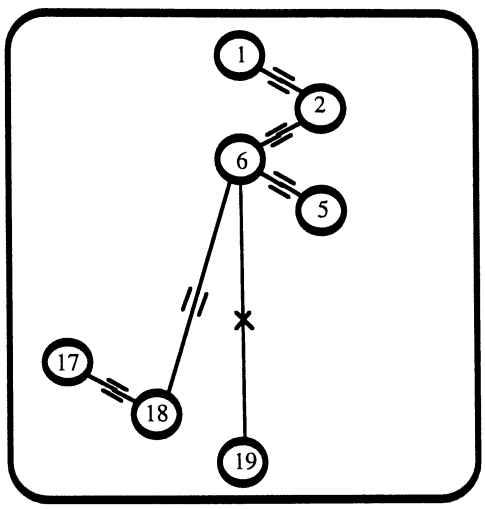

FIG. 7.

We deduce that

$$E[T] \leq n\phi(\rho) + \left\lfloor \frac{\log n}{\log 2} \right\rfloor,$$

and finally

$$\tau_n \leq n\phi(\rho) + \left\lfloor \frac{\log n}{\log 2} \right\rfloor.$$

This ends the proof of Theorem 1. □

COROLLARY 1. *The function ϕ is the largest bounded nonnegative function satisfying* (i) *and* (ii).

When $\rho < 1$, we can obtain a sharper upper bound for τ_n than in Theorem 1, computing the average complexity of our quasi-optimal algorithm by martingale arguments.

COROLLARY 2. *When* $\rho < 1$,

$$\tau_n \leq n\phi(\rho) + C(\rho) \quad \textit{with } C(\rho) = \sum_{k \geq 0} \left(1 - \phi\left(\rho^{2^k}\right)\right).$$

PROOF. Our algorithm avoids comparisons between components having different values until a time $T' = T - Z + 1$: considering the martingale $M_{t \wedge T'}$, we deduce that

$$E[T] = n\phi(\rho) - 1 + \sum_{k \geq 1} \left(1 - \phi\left(\rho^{2^{k-1}}\right)\right) E[h(Y_k)].$$

In order to see that $C(\rho)$ is the sum of a convergent series, we observe that $1 - \phi$ is less than $1/3$, and is $O(x)$ at 0. This last fact results from

$$\frac{1+\rho}{1-\rho} + \frac{1+\rho^2}{2(1-\rho^2)} + \frac{1+\rho^4}{4(1-\rho^4)} + \cdots = 2\frac{1+\rho}{1-\rho} - 2\left(\sum_{n \geq 1} 2^{-|n|_2}\rho^n\right)$$

and thus

$$1 - \phi = \frac{1-\rho}{1+\rho}\left(\sum_{n \geq 1} 2^{-|n|_2}\rho^n\right),$$

in which $|n|_2$ is the exponent of factor 2 in the factor decomposition of n. □

8. A lower bound for μ_n. As Alonso, Reingold and Schott (1994) have already treated the case $\rho = 1$, we treat only the case $\rho < 1$. In this section we prove the following.

PROPOSITION 7.

$$E[T] \geq \frac{1+\rho}{2}\varphi(\rho)n - 1 - o(1).$$

The beginning of the proof is the same as for the composition problem. Using Lemma 1, we obtain again

$$E(T) \geq \varphi(\rho)n - \sum_{k\geq 1} \lambda_k E(C_k(T)) \tag{8.1}$$

but $\sum_{k\geq 1}\lambda_k E(C_k(T))$ is not as easily bounded as in Section 6. Let $R_t = (R_k(t))_{k\geq 1}$ be defined by

$$R_t = C_t - e_{d(C_t)},$$

so that

$$T = \inf\left\{ t \geq 0 \,\middle|\, d(C_t) > \sum_{k=1}^{+\infty} kR_k(t) \text{ or } C_t = e_0 \right\}. \tag{8.2}$$

The key point is the following.

PROPOSITION 8. *We have*

$$\sum_{k=1}^{+\infty} kE(R_k(T)) \leq \frac{(1-\rho)n}{2} + o(1).$$

But we also need the following trite statement.

LEMMA 2. *Let $(\alpha_k)_{1\leq k\leq n}$ and $(\beta_k)_{1\leq k\leq n}$ be two sequences of positive numbers. Then we have*

$$\sum_{k=1}^{n} \alpha_k x_k \leq A \quad \Rightarrow \quad \sum_{k=1}^{n} \beta_k x_k \leq A \max\{\beta_k/\alpha_k \mid 1 \leq k \leq n\}.$$

PROOF OF PROPOSITION 7. Using Lemma 2 and Proposition 8, we obtain that

$$E\left[\sum_{k=1}^{+\infty} \lambda_k R_k(T)\right] \leq \frac{(1-\rho)}{2}\lambda_1 n + o(1),$$

provided that we have

$$\lambda_1 \geq \frac{\lambda_k}{k}. \tag{8.3}$$

This last point follows at once from the fact that $\phi(x)$ is decreasing and takes the values 1, respectively $2/3$, at $x = 0$, respectively 1. Noticing that

$$1 = \|\varphi\|_\infty \geq \sum_{k=1}^{+\infty} \lambda_k C_k(T) - \sum_{k=1}^{+\infty} \lambda_k R_k(T) \geq 0,$$

we deduce

$$E\left[\sum_{k=1}^{+\infty} \lambda_k C_k(T)\right] \leq \frac{(1-\rho)\varphi(\rho)}{2} n + 1 + o(1).$$

Finally (8.1) becomes:

$$E[T] \geq \frac{1+\rho}{2}\varphi(\rho)n - 1 - o(1). \qquad \square$$

For the proof of Proposition 8, and also for the proof of the reverse inequality in Section 9, we need one more lemma. We define X_k by

$$X_k = \begin{cases} 1, & \text{if } x_k \text{ is red,} \\ -1, & \text{if } x_k \text{ is blue,} \end{cases}$$

and we set

$$S_k = \sum_{j=1}^{k} X_j.$$

We shall use a Chernoff bound [see Bollobas (1985), page 12, for instance].

LEMMA 3. *For* $0 < \alpha \leq 1$ *and for any positive* λ,

$$P\left(\left|S_n - n\frac{1-\rho}{1+\rho}\right| \geq \lambda n^{(1+\alpha)/2}\right) \leq \exp\left(-\frac{1}{2}\lambda^2 n^{\alpha}\right).$$

PROOF OF PROPOSITION 8. In the first place we notice that

$$E\left[\sum_{k=1}^{+\infty} kR_k(T)\right] = E\left[\sum_{k=1}^{+\infty} kR_k(T)1_{C_T \neq e_0}\right].$$

Set $N = \#C_T$. Let c_1 be a connected component of the graph G_T with maximal value $d(C_T)$ and let $c_2, c_3, \ldots, c_N$ denote the other useful components at time T. Let us define $\Sigma(j)$ as

$$\Sigma(j) = \sum_{i \in c_j} X_i.$$

The value $v(c_j)$ of a connected component c_j is given by $\nu(c_j) = |\Sigma(j)|$. Thus we have

$$\sum_{i=2}^{N} |\Sigma(i)| = \sum_{k=1}^{+\infty} kR_k(T). \tag{8.4}$$

Let ε_k be the sign of $\Sigma(1)\Sigma(k)$. The stopping condition (8.2), being written

$$|\Sigma(1)| > \sum_{i=2}^{N} |\Sigma(i)|,$$

entails that $\Sigma(1)$ and S_n have the same sign and that

$$|S_n| = |\Sigma(1)| + \sum_{i=2}^{N} \varepsilon_i |\Sigma(i)|.$$

Proposition 2 entails

$$E\left[\varepsilon_k \,\middle|\, |\Sigma(1)| = \nu \text{ and } |\Sigma(k)| = m\right] = \frac{(1-\rho^{\nu})(1-\rho^{m})}{(1+\rho^{\nu})(1+\rho^{m})} \geq \frac{(1-\rho^{\nu})(1-\rho)}{(1+\rho^{\nu})(1+\rho)}.$$

We deduce that

$$E\big[|S_n|\big] \geq E\left[|\Sigma(1)| + \frac{(1-\rho^{|\Sigma(1)|})(1-\rho)}{(1+\rho^{|\Sigma(1)|})(1+\rho)} \sum_{i=2}^{N} |\Sigma(i)|\right]$$

and, from the stopping condition,

$$E\big[|S_n|\big] \geq E\left[\left(1 + \frac{(1-\rho^{|\Sigma(1)|})(1-\rho)}{(1+\rho^{|\Sigma(1)|})(1+\rho)}\right) \sum_{i=2}^{N} |\Sigma(i)|\right]. \tag{8.5}$$

Noticing that

$$|\Sigma(1)| \geq \tfrac{1}{2}|S_n|,$$

we can deduce from Lemma 3 that, for any positive α less than $(1-\rho)/2(1+\rho)$, there exist positive $K(\alpha)$ and $r(\alpha)$ such that

$$P(|\Sigma(1)| \leq \alpha n) \leq K(\alpha) e^{-r(\alpha)n},$$

obtaining a bound for the right-hand side of (8.5):

$$\begin{aligned} &E\left[\left(1 + \frac{(1-\rho^{|\Sigma(1)|})(1-\rho)}{(1+\rho^{|\Sigma(1)|})(1+\rho)}\right) \sum_{i=2}^{N} |\Sigma(i)|\right] \\ &\geq \left(1 + \frac{(1-\rho^{\alpha n})(1-\rho)}{(1+\rho^{\alpha n})(1+\rho)}\right) E\left[\sum_{i=2}^{N} |\Sigma(i)| 1_{|\Sigma(1)| \geq \alpha n}\right] \qquad (8.6) \\ &= \left(\frac{2}{1+\rho} + o\left(\frac{1}{n}\right)\right)\left(E\left[\sum_{i=2}^{N} |\Sigma(i)|\right] + o(1)\right). \end{aligned}$$

Relations (8.4), (8.5) and (8.6) yield that

$$\frac{1+\rho}{2} E\big[|S_n|\big] + o(1) \geq E\left[\sum_{k=1}^{+\infty} k R_k(T)\right].$$

Finally, Proposition 8 follows from a consequence of Lemma 3: we have for $\rho < 1$,

$$E\big[|S_n|\big] = \frac{1-\rho}{1+\rho} n + o(1)$$

since

$$E[S_n] = \frac{1-\rho}{1+\rho} n$$

and

$$\begin{aligned} 0 &\leq E\big[|S_n| - S_n\big] \\ &= 2E\big[|S_n| 1_{S_n < 0}\big] \\ &\leq 2nP(S_n < 0) \\ &\leq 2n \exp\left(-\frac{1}{2}\left(\frac{1-\rho}{1+\rho}\right)^2 n\right). \end{aligned}$$

□

9. A quasi-optimal algorithm for the majority problem. To prove the reverse inequality, we shall describe a quasi-optimal algorithm. In Lemma 2 we have equality, instead of the last inequality, with $A = \sum_{k=1}^{n} \alpha_k x_k$ at the necessary condition that all the x_k are 0 except for the x_i such that β_i/α_i is maximum. Thus relation (8.3) gives a hint of what should be a quasi-optimal algorithm: this relation reflects the fact that at time T, when the stopping condition is first satisfied, it is optimal that each connected component has value 1 (all the x_k are 0 except x_1), with the exception of the connected component with maximum value. Our quasi-optimal algorithm satisfies this condition by computing the composition of a prescribed set A containing approximately $((1 + \rho)/2)n$ elements and leaving the remaining $((1 - \rho)/2)n$ elements untouched (see Proposition 8).

This optimal rule was all but obvious to us: in the uniform case, a quasi-optimal algorithm is obtained by computing the composition of the whole set. Furthermore, comparing the elements of A^c pairwise has a cost, but it has also some advantages. The fall of the sum of the values appearing in the right-hand side of the stopping condition (4.1), resulting from the fact that two connected components with value 1 are sometimes replaced with one connected component with value 0 in this sum, is a progress towards the solution of the majority problem. Relation (8.3) proves implicitly that the cost of this progress is too high: comparing two elements of A^c, we run the risk that some x_i other than x_1 is positive.

Consider the following algorithm: set

$$N(\alpha) = \left\lfloor \frac{1 + \rho}{2} n + n^{\alpha} \right\rfloor,$$

$\alpha > 1/2$. Then apply the quasi-optimal algorithm of Section 7 to determine the composition of the set $\{x_k \mid 1 \le k \le N(\alpha)\}$, at an average cost less than or equal to

$$N(\alpha)\phi(\rho) + \left\lfloor \frac{\log N(\alpha)}{\log 2} \right\rfloor.$$

At this stage we obtain a connected component with value $|S_{N(\alpha)}|$, some useless components with value 0, and exactly $n - N(\alpha)$ components with value 1. As a consequence, a sufficient condition for the majority to be known at this stage is

$$|S_{N(\alpha)}| > n - N(\alpha).$$

If $|S_{N(\alpha)}| \le n - N(\alpha)$, we compute the composition of the whole set, at a cost T less than or equal to n. Proposition 8 entails that $P(|S_{N(\alpha)}| \le n - N(\alpha))$ is exponentially small in a power of n, and so we deduce that

$$\begin{aligned} \mu_n &\le N(\alpha)\phi(\rho) + \left\lfloor \frac{\log N(\alpha)}{\log 2} \right\rfloor + o(1) \\ &\le n \frac{1 + \rho}{2} \phi(\rho) + O(n^{\alpha}) \end{aligned}$$

10. A quasi-optimal algorithm for unknown probabilities. It is natural to ask for a quasi-optimal algorithm when ρ is unknown. The quasi-optimal algorithm for the composition problem does not depend on ρ, but it does for the majority problem, through the prescribed set A containing approximately $((1+\rho)/2)n$ elements. However the first round of this algorithm is a sequence of at least $n/4$ [and actually $((1+\rho)/4)n$] Bernoulli trials, each of these trials having a probability of success

$$p_+(1,1) = \frac{1+\rho^2}{(1+\rho)^2}.$$

For n large, the first $n/4$ trials give a pretty good estimate of $p_+(1,1)$, and thus of ρ. One easily deduces a quasi-optimal algorithm for ρ unknown.

11. Concluding remarks. Very likely, the upper bound in Theorem 2 can be improved, because the size $N(\alpha)$ of the prescribed set of Section 9 is too large. The smallest possible size of the prescribed set $A_k = \{1, 2, 3, \ldots, k\}$ on which we should work would be T_n, defined by

$$T_n = n \wedge \inf\{k \geq 0 \,|\, |S_k| > n - k\}.$$

It turns out, by the very definition of T_n, that the stopping condition for the majority problem is satisfied just at the moment when we discover the composition of A_{T_n}, through the optimal algorithm for the composition problem. Now we have easily the following proposition.

PROPOSITION 9. *If* $\rho = 1$,

$$E[T_n] = n - \sqrt{\frac{2n}{\pi}} + O(1),$$

while if $\rho < 1$

$$E[T_n] = (1+\rho)\left\lceil \frac{n+1}{2} \right\rceil + o(1)$$
$$= \frac{1+\rho}{2} n + O(1).$$

Thus we can expect, when we work on the prescribed set $A = \{1, 2, 3, \ldots, T_n\}$, to pay an average cost bounded by

$$E[\phi(1)T_n + O(\log T_n)] \approx \phi(1)\left(n - \sqrt{\frac{2n}{\pi}} + O(1)\right) + O(\log E[T_n])$$
$$\approx \frac{2}{3}\left(n - \sqrt{\frac{2n}{\pi}}\right) + O(\log n)$$

comparisons if $\rho = 1$, and similarly

$$E[\phi(\rho)T_n + O(1)] \approx \phi(\rho)\frac{1+\rho}{2} n + O(1)$$

comparisons if $\rho < 1$. The first estimate is just the result obtained by Alonso, Reingold and Schott (1994), while the second estimate would be a lot better than Theorem 2. We noticed the same tight relation between the asymptotics of $E[T_n]$ and the asymptotic of the optimal cost in the majority problem in a later work of Alonso, Chassaing and Schott (1996) about coin weighing. Of course we are cheating in the previous estimates in many respects, for instance because it is likely that the colors of the elements in A are not conditionally independent given $T_n = k$, and thus the average cost of the algorithm, knowing that $T_n = k$, could be different from $\phi(\rho)k + O(1)$. A more serious problem arises because in order to determine the composition of the prescribed set $A = \{1, 2, 3, \ldots, T_n\}$ using the quasi-optimal algorithm, we need to know T_n when we start the algorithm, while we cannot actually know the value of T_n until very late in the runtime of the algorithm. I believe that these difficulties can be overcome at a low cost (maybe by a more adaptive choice of A, even if it entails sometimes breaking the rule "compare components with the same value"), but a lot of tedious work would be needed to describe and to analyze such an algorithm.

Classical tools from control of stochastic processes are seldom used in proofs of average case optimality of algorithms: other examples can be found in Chassaing (1993) and Alonso, Chassaing and Schott (1996). A survey of celebrated problems in this area, such as the sorting problem and the selection problem, is given in Knuth (1973). Maybe some of the many open problems of average case optimality could be tackled successfully with the help of these powerful tools.

Acknowledgments. I am pleased to acknowledge Laurent Alonso, E. R. Reingold and René Schott for giving me an easy access to an early version of their second paper. I especially thank Laurent Alonso for his help in the proof of Proposition 6, and also an anonymous referee whose detailed comments helped considerably to improve the first version of this paper.

REFERENCES

ALONSO, L., CHASSAING, P. and SCHOTT, R. (1996). A coin-weighing problem. *Random Structures Algorithms* **9** 1–14.

ALONSO, L., REINGOLD, E. R. and SCHOTT, R. (1993). Determining the majority. *Inform. Process. Lett.* **47** 253–255.

ALONSO, L., REINGOLD, E. R. and SCHOTT, R. (1994). The average-case complexity of determining the majority. *SIAM J. Comput.* To appear.

ANDREWS, G. E. (1976). The theory of partitions. In *Encyclopedia of Mathematics* **2**. Addison-Wesley, London.

BERTSEKAS, D. P. (1987). *Dynamic Programming: Deterministic and Stochastic Models*. Prentice-Hall, Englewood Cliffs, NJ.

BOLLOBAS, B. (1985). *Random Graphs*. Academic Press, London.

CHASSAING, P. (1993). Optimality of move-to-front for self-organizing data structures with locality of references. *Ann. Appl. Probab.* **3** 1219–1240.

CUNTO, W. and MUNRO, J. I. (1984). Average case selection. In *Proceedings of the 16th Annual ACM Symposium on Theory of Computing* 369–375.

HAKIMI, S. L. and SCHMEICHEL, E. F. (1984). An adaptive algorithm for system level diagnosis. *J. Algorithms* **5** 526–530.

HARDY, G. H. and RAMANUJAN, S. (1918). Asymptotic formulae in combinatory analysis. In *Collected Papers of S. Ramanujan* 276–309. Chelsea, New York (1962).

KNUTH, D. E. (1973). *The Art of Computer Programming* **3**: *Sorting and Searching*. Addison-Wesley, Reading, MA.

PREPARATA, F., METZE, G. and CHIEN, R. T. (1967). On the connection assignment problem of diagnosable systems. *IEEE Trans. Comput.* **16** 848–854.

SAKS, M. R. and WERMAN, M. (1991). On computing the majority by comparisons. *Combinatorica* **11** 383–387.

SCHMEICHEL, E., HAKIMI, S. L., OTSUKA, M. and SULLIVAN, G. (1990). A parallel fault identification algorithm. *J. Algorithms* **11** 231–241.

INSTITUT ELIE CARTAN, URA CNRS 750
UNIVERSITÉ HENRI POINCARÉ–NANCY I
BP 239
54506 VANDŒUVRE-LÈS-NANCY
FRANCE
E-MAIL: chassain@iecn.u-nancy.fr