Combinatorics, Probability and Computing (2005) 14, 17–24.
DOI: 10.1017/S0963548304006704 Printed in the United Kingdom

The k-Equal Problem

MARTIN AIGNER

Freie Universität Berlin, Fachbereich Mathematik und Informatik
Arnimallee 14, D-14195 Berlin
(e-mail: aigner@math.fu-berlin.de)

Received 30 December 2002; revised 16 June 2003

Suppose we are given n coloured balls and an integer k between 2 and n. How many colour-comparisons $Q(n,k)$ are needed to decide whether k balls have the same colour? The corresponding problem when there is an (unknown) linear order with repetitions on the balls was solved asymptotically by Björner, Lovász and Yao, the complexity being $\theta(n \log \frac{2n}{k})$. Here we give the exact answer for $k > \frac{n}{2}$: $Q(n,k) = 2n - k - 1$, and the order of magnitude for arbitrary k : $Q(n,k) = \theta(\frac{n^2}{k})$.

1. Introduction

Suppose we are given n coloured balls and an integer k, $2 \leqslant k \leqslant n$. Two players, Paul and Carole, play the following game. At any stage Paul chooses two balls and asks whether they have the same colour, whereupon Carole answers 'yes' or 'no'. The game ends when Paul either produces k balls coloured alike or states that no k balls have the same colour. How many questions $Q(n,k)$ are needed in the worst case?

Two variants of this problem have been studied in detail. Suppose the n balls are coloured with two colours, and Paul has to produce a majority ball (or state that there is no majority). This variant is known as the majority problem and was solved first by Saks and Werman [9], and later by Alonso, Reingold, Schott [4, 5] and Wiener [10] using different methods. They answer it exactly: $n - B(n)$ comparisons are needed, where $B(n)$ is the number of 1s in the binary representation of n. When there is an arbitrary number of colours the solution $\lceil \frac{3n}{2} \rceil - 2$ was given by Fisher and Salzberg [8]. A general account of this and related results is contained in Aigner [2, 3].

The other variant was considered by Björner, Lovász and Yao. Suppose we are given n real numbers (or equivalently a linear order on the balls with repetitions). How many

comparisons are necessary to decide whether k of the numbers are equal? They give the asymptotic answer $\theta(n \log \frac{2n}{k})$.

In our situation we are at the other end. The coloured balls may be thought of as an antichain with repetitions. The general problem would then consist of an arbitrary poset (with repetitions). For example, for a linear order the case $k = 2$ amounts to sorting the n numbers, while for an antichain we clearly have $Q(n, 2) = \binom{n}{2}$. For a general poset P, the case $k = 2$ amounts to producing the poset P; see, *e.g.*, Aigner [1, Chapter 5].

Let us return to the topic of this note. It is to be expected that the cases $k > \frac{n}{2}$ and $k \leqslant \frac{n}{2}$ are of an entirely different nature. In the first case, a k-set of equally coloured balls is unique, while for $k \leqslant \frac{n}{2}$ uniqueness is no longer guaranteed. This is reflected in the following two results.

Theorem 1.1. *Let $k > \frac{n}{2}$. Then $Q(n,k) = 2n - k - 1$.*

Theorem 1.2. *For general k, $2 \leqslant k \leqslant n$, we have $Q(n,k) = \theta(\frac{n^2}{k})$, more precisely, $\frac{1}{4}\frac{n^2}{k} < Q(n,k) < 2\frac{n^2}{k}$.*

2. Proof of Theorem 1.1

We use an argument similar to that in [8]. Let us look at the upper bound $Q(n,k) \leqslant 2n - k - 1$ first. Paul uses the following algorithm.

Phase 1. Arrange the balls $B_1, B_2, \ldots, B_n$ in linear order, and compare the balls B_i one after the other. We set up a dynamic list L and a reservoir R. Initially, $L = \{B_1\}$, $R = \varnothing$. Suppose that before B_j is compared the list is $L = C_1 C_2 \ldots C_s$ with reservoir R. Now compare B_j to the last ball C_s of the list. If they have the same colour, put B_j into R. Otherwise, enlarge L by moving B_j to the end and putting a ball Z of R behind B_j (in case $R \neq \varnothing$). Phase 1 ends when the last ball B_n has been handled accordingly, so we have made $n - 1$ comparisons so far.

The following facts are immediate.

(a) All balls in R have the same colour which is equal to the colour of the last ball of L.
(b) Neighbouring balls in L have different colours.
(c) Since $k > \frac{n}{2}$, the only possible k-set of equally coloured balls contains the last ball of L.

Phase 2. Let $L = D_t D_{t-1} \ldots D_1 B A$ be the list with $|R| = r$, and thus $t = n - 2 - r$. Let us call the colour of A the *majority colour*, and all balls of this colour *majority balls*. Any ball with a colour different from A is called a *non-majority ball*, and we know that B is such a ball. If at the outset of Phase 2, either $r + 1 \geqslant k$ or $n + r + 2 - 2k \leqslant 0$, then we are done. In the first case, we have found k majority balls, and in the second case there is no such set since the number of majority balls is at most

$$r + \left\lceil \frac{n-r}{2} \right\rceil = \left\lceil \frac{n+r}{2} \right\rceil \leqslant \frac{2k-2}{2} = k - 1.$$

https://doi.org/10.1017/S0963548304006704 Published online by Cambridge University Press

Now we scan the list L from the end, first comparing D_1 with A, then D_2 with A, and so on. Note that whenever D_i is a majority ball ($i \geqslant 1$), then D_{i+1} is a non-majority ball. Suppose that up to a certain stage we have received ℓ answers 'yes' and m answers 'no'. The game is over when, for the first time, either after a 'yes' answer,

$$\ell + r + 1 \geqslant k, \tag{2.1}$$

or after a 'no' answer,

$$m + \ell + 1 + \left\lfloor \frac{n - r - 2 - 2\ell - m}{2} \right\rfloor \geqslant n - k + 1, \tag{2.2}$$

where ℓ accounts for non-majority balls after each 'yes' answer, 1 for B, and the last summand for the remaining balls. Inequality (2.2) is equivalent to

$$m \geqslant n + r - 2k + 2. \tag{2.3}$$

It remains to estimate the number of questions $\ell + m$. Suppose (2.1) occurs first. Then $\ell + r + 1 = k$ and $2\ell - 1 + m \leqslant n - r - 2$ (note that the ℓth 'yes' may occur when D_t is compared to A). Hence we obtain

$$\ell + m \leqslant n - r - 1 - \ell = n - k.$$

Suppose that (2.3) occurs first. Then $\ell \leqslant k - 2 - r$ by (2.1) and $m = n + r - 2k + 2$ by (2.3). Adding, we obtain again

$$\ell + m \leqslant n - k.$$

Hence Phase 2 uses at most $n - k$ questions, and so $Q(n,k) \leqslant 2n - k - 1$.

We come to the lower bound $Q(n,k) \geqslant 2n - k - 1$. Carole sets up the following data structure. She constructs a dynamic graph G on the balls where the edges correspond to the 'no' answers. The balls are dynamically labelled A, B or C with weights w, where all balls labelled B have weight $1/2$ and all balls labelled C have weight 0. Let $\mathscr{A}$, $\mathscr{B}$, $\mathscr{C}$ be the sets of balls with labels A, B and C, respectively, and set $\sum = \sum_{X \in \mathscr{A} \cup \mathscr{B}} w(X)$. At the start all balls are labelled A with weight 1, thus $\sum = n$. If $X \in \mathscr{A}$ has weight i, then we write $X = A(i)$.

Now Carole answers according to the following rules, where we keep in mind that the A-balls aim for the k-majority.

(I′) $A(i) : A(j)$ 'yes' if $\sum = k$, and the two balls are merged into a single ball $A(i+j)$.
'no' if $\sum > k$, and an edge is inserted. In this case

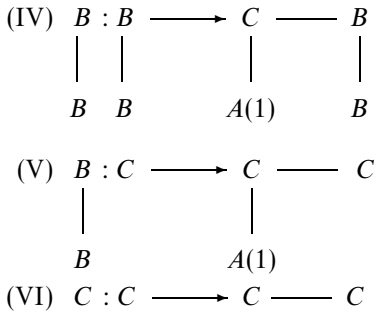
(I″) $i = j = 1$ and we make the label change $A(1) : A(1) \longrightarrow B - B$.

In all other cases the answer is 'no', with the following label changes:

(II) $A(i) : B \longrightarrow A(i) \longrightarrow C$, with B and $A(1)$ attached

(III) $A(i) : C \longrightarrow A(i) \longrightarrow C$

https://doi.org/10.1017/S0963548304006704 Published online by Cambridge University Press



We note the following easy facts.

(a) An edge of B-balls is only created when $A(1) : A(1)$ are compared.
(b) The B-balls constitute a matching, and there is no edge between different B-edges, between $A(i)$ and $A(j)$, and between $A(i)$ and B.
(c) Σ is only reduced (by 1) in case (I″) and remains unchanged otherwise; Σ never drops below k. Furthermore, as long as $\Sigma > k$, all balls in $\mathscr{A}$ have weight 1. We have $|\mathscr{A}| \geqslant 1$ at any time, since $\sum_{Y \in \mathscr{B}} w(Y) \leqslant \frac{n}{2} < k$.
(d) At a certain stage the number of 'yes' answers is $\sum_{X \in \mathscr{A}} w(X) - |\mathscr{A}|$. This follows from (I′). The number of 'no' answers (= edges) is at least $2|\mathscr{C}| + \sum_{Y \in \mathscr{B}} w(Y)$. This follows from (I″) to (VI).
(e) At any stage a k-majority is possible. Indeed, since there are no edges between $A : A$, $A : B$ and different B-edges, all weighted balls in $\mathscr{A}$ and one from each edge in $\mathscr{B}$ may be coloured alike, summing to $\sum \geqslant k$.

Claim. As long as $|\mathscr{A}| + |\mathscr{B}| \geqslant 2$, there is also a non-majority possible, and hence the game is not over.

Case 1: $\Sigma > k$. Then all balls in $\mathscr{A}$ have weight 1, and so all n balls may have different colours.

Case 2: $\Sigma = k$. If $|\mathscr{B}| \geqslant 2$, then $\sum_{X \in \mathscr{A}} w(X) < k$, and thus a non-majority is possible by assigning different colours to different A-balls and to all other balls. Suppose then $\mathscr{B} = \varnothing$ and $|\mathscr{A}| \geqslant 2$. But here the same argument works since $w(X) < k$ for all $X \in \mathscr{A}$.

Hence we conclude that the game is only over when $|\mathscr{A}| = 1$, $\mathscr{B} = \varnothing$, $\Sigma = k$, and thus $|\mathscr{C}| = n - k$. Invoking (d) we find that Paul has asked at least

$$k - 1 + 2(n - k) = 2n - k - 1$$

questions, and the proof is complete.

Remarks. **(1)** When k drops below $\frac{n}{2}$, the situation changes completely. Let n be even; then we have just seen that $Q(n, \frac{n}{2} + 1) = \frac{3n}{2} - 2$. By an argument similar to that of the proof of the theorem, it can be shown that $Q(n, \frac{n}{2}) = 2n - 3$ for $n \geqslant 6$.

https://doi.org/10.1017/S0963548304006704 Published online by Cambridge University Press

(2) In a variant which seems to favour Paul, Carole is only allowed c colours (unknown to Paul). It can be shown that in the strategy of Carole used in Theorem 1.1 she can force Paul to ask $2n-k-1$ questions using no more than 5 colours, for any n.

3. Proof of Theorem 1.2

We may restrict ourselves to $k \leqslant \frac{n}{2}$. Here the lower bound is easier. Carole always answers 'no'. Let G be the dynamic graph on the n balls with edges corresponding to the comparisons. A k-set of identically coloured balls is possible as long as the independence number $\alpha(G)$ is at least k. By Turán's Theorem the minimal number of edges such that $\alpha(G) \leqslant k-1$ is achieved when G is the union of $k-1$ complete subgraphs of nearly equal size. Hence

$$Q(n,k) \geqslant (k-1)\binom{\frac{n}{k-1}}{2} = \frac{1}{2}\left(\frac{n^2}{k-1} - n\right) > \frac{1}{4}\frac{n^2}{k},$$

where the last inequality follows from $2k \leqslant n$.

Now to the upper bound. Let $n = qk + r$, $0 \leqslant r < k$. Paul arranges the balls $1, 2, \ldots, n$ around a circle in clockwise fashion, and proceeds in $k-1$ rounds. In the first round he compares i with $i+1, \ldots, i+q \pmod n$ for all i. This takes qn comparisons. The 'yes' answers partition $\{1, \ldots, n\}$ into sets $A_1, \ldots, A_m$, where balls in the same set have the same colour. Let us set $\mathscr{C}(1) = \{A_1, \ldots, A_m\}$, with $\sum_{i=1}^{m} |A_i| = n$. Any set $A \in \mathscr{C}(1)$ is an 'arc' $A = \{a_1, \ldots, a_s\}$ read clockwise with $a_{i+1} - a_i \leqslant q$. Now either $s \geqslant k$, in which case we are finished, or the q balls following a_s have different colours from A. Furthermore, any ball $b \notin A$ in the arc, that is, $a_1 < b < a_s$, also has a different colour. We say that A *dominates* at least q balls. Suppose that all sets $A \in \mathscr{C}(1)$ have size $\leqslant k-1$.

Claim. If the sets $A_{i_1}, \ldots, A_{i_\ell}$ in $\mathscr{C}(1)$ form a possible $\geqslant k$-set with equal colour, then at least one of the sets contains at least two elements.

Suppose otherwise with $A_{i_j} = \{a_j\}$, $j = 1, \ldots, k$. Then by what we have just seen, any two consecutive balls are separated by at least q balls around the circle. It follows that $n \geqslant kq + k$, which contradicts $r < k$.

So, for the next round we need only compare sets $A \in \mathscr{C}(1)$ with $|A| \geqslant 2$ in clockwise fashion. In every round we compare A with the next q undecided balls. Let us fix some notation. After round i $(1 \leqslant i \leqslant k-1)$, $\mathscr{C}(i) = \{A_1, \ldots, A_m\}$ is the partition into equally coloured sets determined by the 'yes' answers. We denote by $s(A)$ the smallest number such that $A \in \mathscr{C}(i)$ dominates at least $s(A)q$ balls of different colour; $\#A$ denotes the number of comparisons which involved balls of A in rounds 2 to i.

Now either there is $A \in \mathscr{C}(i)$ with $|A| \geqslant k$, in which case we are finished, or we assume inductively that the following holds:

(a) $|A| \leqslant i \Longrightarrow s(A) \geqslant |A|, \quad \#A \leqslant (|A|-1)q,$

(b) $|A| \geqslant i+1 \Longrightarrow s(A) \geqslant i, \quad \#A \leqslant [\min(|A|-2, s(A)-1)]q.$

Note that (a) and (b) hold for $i = 1$.

Claim. If, after round i, the sets $A_1, \ldots, A_\ell$ of $\mathscr{C}(i)$ constitute a possible $\geqslant k$-set of equally coloured balls, then at least one of them contains at least $i+1$ balls.

https://doi.org/10.1017/S0963548304006704 Published online by Cambridge University Press

Suppose not, and set $f_t = \#\{j : |A_j| = t\}$, $1 \leqslant t \leqslant i$. Then $\sum_{t=1}^{i} tf_t = |\bigcup_{j=1}^{\ell} A_j| \geqslant k$. Invoking (a) we see that a t-set A is separated by at least tq balls from the next set, whence

$$n \geqslant \sum_{t=1}^{i} tf_t + \sum_{t=1}^{i} (tf_t)q \geqslant kq + k,$$

which cannot be.

Hence Paul need only compare sets $A \in \mathscr{C}(i)$ with $|A| \geqslant i+1$ in round $i+1$, and in fact he only compares all sets $A \in \mathscr{C}(i)$ with

$$|A| \geqslant i+1 \quad \text{and} \quad s(A) = i.$$

We have to check the conditions (a) and (b) after round $i+1$. Clearly, any set $B \in \mathscr{C}(i+1)$ is a union of sets $A_1, \ldots, A_h$ of $\mathscr{C}(i)$.

Case 1: $h = 1$, that is, $B = A \in \mathscr{C}(i)$. If $|A| \leqslant i$ or $|A| \geqslant i+1$ and $s(A) \geqslant i+1$, then A has not been compared, $s(B) = s(A)$, and (a) and (b) are satisfied. If $|A| = i+1$, $s(A) = i$, then A has been compared (with all q balls following A of different colour), hence $s(B) = s(A) + 1$. Thus $s(B) = i+1 = |B|$ and $\#B = \#A + q \leqslant (|A| - 2)q + q = (|B| - 1)q$, and (a) holds. Finally, if $|A| \geqslant i+2$, $s(A) = i$, then $s(B) = s(A) + 1 = i+1$. Furthermore, $\#A \leqslant [\min(|A| - 2, s(A) - 1]q = (i-1)q$, thus $\#B = \#A + q \leqslant iq \leqslant [\min(|B| - 2, s(B) - 1]q$, since $|B| - 2 \geqslant i$ and $s(B) - 1 = i$.

Case 2: $B = A_1 \cup \cdots \cup A_h$ in clockwise order, $h \geqslant 2$. Since only $A \in \mathscr{C}(i)$ with $|A| \geqslant i+1$, $s(A) = i$ are compared, we have

$$|A_j| \geqslant i+1,\ s(A_j) = i \quad \text{for} \quad j = 1, \ldots, h-1,$$
$$|A_h| \geqslant i+1,\ s(A_h) \geqslant i \quad \text{or} \quad |A_h| \leqslant i,\ s(A_h) \geqslant |A_h|.$$

It follows that

$$s(B) \geqslant (h-1)i + \min(i, |A_h|) \geqslant i+1.$$

If A_h is also compared, that is, $|A_h| \geqslant i+1$, $s(A_h) = i$, then

$$\begin{aligned} \#B = \sum_{j=1}^{h} \#A_j + hq &\leqslant \sum_{j=1}^{h} [\min(|A_j| - 2, i-1)]q + hq \\ &= (i-1)hq + hq = ihq \\ &\leqslant [\min(|B| - 2, s(B) - 1)]q, \end{aligned}$$

since $|B| \geqslant (i+1)h$, and hence $|B| - 2 \geqslant (i+1)h - 2 \geqslant ih$ because of $h \geqslant 2$, and $s(B) \geqslant i(h-1) + (i+1) = ih + 1$, and thus $s(B) - 1 \geqslant ih$.

https://doi.org/10.1017/S0963548304006704 Published online by Cambridge University Press

Finally, if A_h is not compared, then either $|A_h| \geqslant i+1$, $s(A_h) \geqslant i+1$ or $|A_h| \leqslant i$, $s(A_h) \geqslant |A_h|$. This gives

$$\begin{aligned}\#B = \sum_{j=1}^{h} \#A_j + (h-1)q &\leqslant \sum_{j=1}^{h-1} [\min(|A_j|-2, i-1)]q \\ &\quad + \begin{cases} [\min(|A_h|-2, s(A_h)-1)]q \\ (|A_h|-1)q \end{cases} \quad + (h-1)q \\ &= i(h-1)q + \begin{cases} [\min(|A_h|-2, s(A_h)-1)]q \\ (|A_h|-1)q \end{cases} \\ &\leqslant [\min(|B|-2, s(B)-1)]q,\end{aligned}$$

since $|B| \geqslant (i+1)(h-1) + |A_h|$. Hence, with $h \geqslant 2$,

$$|B| - 2 \geqslant i(h-1) + h - 3 + |A_h| \geqslant i(h-1) + |A_h| - 1,$$

and

$$s(B) \geqslant i(h-1) + \begin{cases} s(A_h), \\ |A_h|, \end{cases}$$

and thus

$$s(B) - 1 \geqslant i(h-1) + \begin{cases} s(A_h) - 1, \\ |A_h| - 1. \end{cases}$$

Performing rounds 2 to $k-1$, we see that either we obtain a $\geqslant k$-set of equal colours, or that at the latest after round $k-1$ no such set is possible. Thus the game is finished after round $k-1$. Counting the number of comparisons in rounds 2 to $k-1$ we find that after the last round $\ell \leqslant k-1$

$$\sum_{A \in \mathscr{C}(\ell)} \#A \leqslant \sum_{A \in \mathscr{C}(\ell)} (|A|-1)q < nq.$$

Altogether, this gives, with round 1, at most $2nq \leqslant 2\frac{n^2}{k}$ comparisons, and the proof is complete. □

References

[1] Aigner, M. (1988) *Combinatorial Search*, Wiley.
[2] Aigner, M. (2004) Variants of the majority problem. *Discrete Math.* **137** 3–25.
[3] Aigner, M. Two colors and more. Preprint.
[4] Alonso, L., Reingold, E. and Schott, R. (1993) Determining the majority. *Inform. Process. Lett.* **47** 253–255.
[5] Alonso, L., Reingold, E. and Schott, R. (1997) Average-case complexity of determining the majority. *SIAM J. Comput.* **26** 1–14.

https://doi.org/10.1017/S0963548304006704 Published online by Cambridge University Press

[6] Björner, A., Lovász, L. and Yao, A. (1992) Linear decision trees: Volume estimates and topological bounds. In *Proc. 24th ACM Symp. on Theory of Computing*, ACM Press, New York, pp. 170–177,
[7] Du, D. and Hwang, F. (1993) *Combinatorial Group Testing*, World Scientific.
[8] Fisher, M. and Salzberg, S. (1982) Finding a majority among *n* votes. *J. Algorithms* **3** 375–379.
[9] Saks, M. and Werman, M. (1991) On computing majority by comparisons. *Combinatorica* **11** 383–387.
[10] Wiener, G. (2002) Search for a majority element. *J. Statist. Plann. Inference* **100** 313–318.

https://doi.org/10.1017/S0963548304006704 Published online by Cambridge University Press