#16

J. Appl. Prob. 16, 657–661 (1979)
Printed in Israel

RANDOM EXCHANGES OF INFORMATION

DAVID W. BOYD* AND
J. MICHAEL STEELE,** *The University of British Columbia*

Abstract

Suppose that n persons each know a different piece of information, and that whenever a pair of persons talk on the telephone each tells the other all the information that he knows at the time. If calls are made at random, we show that the expected number of calls necessary until everyone knows all n pieces of information is asymptotically 1.5 n log n. This sharpens an earlier result of J. W. Moon.

CONTAGION; RANDOM MATRIX; TELEPHONE PROBLEM

1. Introduction

The object of this paper is to give a solution to a problem raised by Moon [4] concerning a model of random contagion. Of the several ways to describe this model one has become traditional. Suppose that n persons each know a different piece of information, and whenever two of them talk on the telephone each tells the other all the information he knows at the time. The way in which information spreads through such a system suggests a variety of problems.

The popular 'telephone problem', attributed to A. V. Boyd in [4] and [5], is to determine the minimum number of calls $c(n)$ required until everyone has learned all n pieces of information. It is clear that $c(1) = 0$, $c(2) = 1$ and $c(3) = 3$, and A. V. Boyd exhibited a scheme of calls to show that if $n \geqq 4$, $c(n) \leqq 2n - 4$ (such a scheme is described in [5]). He conjectured that $c(n) = 2n - 4$ for $n \geqq 4$, and this was proved independently by Tijdeman [5], Bumby and Spencer [1], and Hajnal, Milner and Szemeredi [3].

Moon [4] considered instead the situation in which the calls are made at random (i.e. the operator chooses two different persons at random from the n and places the call). Suppose C calls are made before each person knows all the

Received 11 May 1978; revision received 25 September 1978.

* Postal address: Department of Mathematics, The University of British Columbia, 2075 Wesbrook Mall, Vancouver, B.C. Canada V6T 1W5.

**Present address: Department of Statistics, Stanford University, Stanford, CA 94305, U.S.A.

Research supported in part by the National Research Council of Canada.

information. The problem is to determine $E(C)$, the expected value of C. Moon gave an elegant proof of the bounds

$$(1-\varepsilon)n \log n \leqq E(C) \leqq (2+\varepsilon)n(\log n)^2,$$

where $\varepsilon > 0$ and n is sufficiently large. Here we will prove the following more precise result.

Theorem.

$$E(C) = \frac{3}{2} n \log n + O(n(\log n)^{1/2}).$$

A different formulation of the telephone problem, attributed to Wirsing in [5], may serve to put our result in a more general context. Let B_{ij} be the $n \times n$ matrix with all entries 0 except for a 1 in the (i,j) and (j,i) positions. Let $A(0) = I$ and $A(t) = A(t-1)(I + B_{ij})$ if there is a call between persons i and j at time t. Then $a_{km}(t) > 0$ if and only if the mth person knows the kth piece of information at time t. One sees this by induction, after observing that $A(t)$ is obtained from $A(t-1)$ by replacing the ith and jth columns by their sum, leaving all other columns unchanged. Thus, if $m = i$ or j, then $a_{km}(t) > 0$ if and only if at least one of $a_{ki}(t-1) > 0$ or $a_{kj}(t-1) > 0$ holds, while if $m \neq i, j$ then $a_{km}(t) = a_{km}(t-1)$. This corresponds to the fact that i and j know the kth piece of information at time t if and only if one of them knew it at time $t-1$. Thus, if the B_{ij} are chosen at random, C is the first time at which $A(t)$ is a strictly positive matrix.

2. The lower bound

Let T_i be the first time at which everyone knows person i's original information, and let T_i^* be the first time at which person i knows everyone's information. In terms of the matrix formulation, T_i is the first time that the ith row of $A(t)$ has no zeros, and T_i^* is the first time that the ith column of $A(t)$ has no zeros. Since each B_{ij} is symmetric, it is clear that one has the following result.

Lemma. The distributions of the vectors $(T_1, \cdots, T_n)$ and $(T_1^*, \cdots, T_n^*)$ are equal.

By definition, $C = \max_{1 \leqq i \leqq n} T_i = \max_{1 \leqq i \leqq n} T_i^*$. Moon used the observation that $E(C) \geqq E(T_1)$ to obtain his lower bound. To improve this bound, it seems necessary to consider all the T_i. Let $T_{(i)}^*$ be the ith largest of the T_i^* so that $T_{(1)}^* \leqq \cdots \leqq T_{(n)}^* = C$. Now let R_i be the number of calls until each of the persons $1, 2, \cdots, i$ has received a call. Then, for $i \leqq n-2$

$$E(C) \geqq E(T_{(i)}^*) + E(R_{n-i-1}). \tag{1}$$

This follows from the fact that, at time $T_{(i)}^*$ either i or $i+1$ persons know everything (since a call involves two callers). For the remaining $n-i-1$ persons to learn everything, each must receive at least one call. If one defines $E(R_0) = E(R_{-1}) = 0$, then (1) is obviously valid for $i = n-1$ and $i = n$. Averaging (1) over i, we have

$$(2) \qquad E(C) \geqq n^{-1}\sum_{i=1}^{n} E(T^*_{(i)}) + n^{-1}\sum_{i=1}^{n-2} E(R_i).$$

But, from the lemma,

$$(3) \qquad \sum_{i=1}^{n} E(T^*_{(i)}) = \sum_{i=1}^{n} E(T^*_i) = \sum_{i=1}^{n} E(T_i) = nE(T_1).$$

Combining (2) and (3) gives

$$(4) \qquad E(C) \geqq E(T_1) + n^{-1}\sum_{i=1}^{n-2} E(R_i).$$

As in [4], $E(T_1)$ can be calculated by observing that $T_1 = X_1 + \cdots + X_{n-1}$, where X_i is the number of calls after i persons know person 1's information until $i+1$ know it. The X_i are independent and geometrically distributed with parameter $p_i = (2i(n-i))/(n(n-1))$. Hence $E(X_i) = 1/p_i$ and

$$(5) \qquad E(T_1) = \sum_{i=1}^{n-1} p_i^{-1} = \frac{n-1}{2}\sum_{i=1}^{n-1}\left\{\frac{1}{i} + \frac{1}{n-i}\right\} = n\log n + O(n).$$

The calculation of $E(R_i)$ is slightly more complicated. We observe that the process of generating calls at random is equivalent to the following: first generate numbers $N_1, N_2, \cdots$ at random from $\{1, 2, \cdots, n\}$. Form the sequence of pairs $(N_1, N_2), (N_3, N_4), \cdots$. Delete from this sequence any *pair* for which $N_{2k-1} = N_{2k}$, obtaining a sequence $(N'_1, N'_2), \cdots$. Then a call is made at time t between N'_{2t-1} and N'_{2t}. Thus R_i is the time at which $N'_1, \cdots, N'_{2t}$ first contains all of $\{1, \cdots, i\}$. Let S be the first time at which $N_1, \cdots, N_S$ contains all of $\{1, \cdots, i\}$ and M be the first time at which $N_1, \cdots, N_{2M}$ contains $\{1, \cdots, i\}$. Clearly $M \geqq S/2$. Next let A_k be 1 if $N_{2k-1} = N_{2k}$ and 0 otherwise. Put $K = \Sigma_{k=1}^{M} A_k$ and note that $R_i \geqq M - K$. Since $E(A_k) = 1/n$, Wald's lemma ([2], p. 380), shows $E(K) = E(M)/n$, so

$$(6) \qquad E(R_i) \geqq (1 - n^{-1})E(M) \geqq (1 - n^{-1})E(S)/2.$$

But $S = B_1 + \cdots + B_i$ where B_j is the time between the occurrence of the $(j-1)$th and jth elements of $\{1, 2, \cdots, i\}$ in $N_1, \cdots, N_S$. So B_j is distributed geometrically with parameter $(i-j+1)/n$. Hence $E(B_j) = n/(i-j+1)$. So $E(S) = n\ \Sigma_{j=1}^{i} j^{-1}$. Combining this with (6) gives

$$(7) \qquad n^{-1}\sum_{i=1}^{n-2} E(R_i) \geqq \frac{1}{2} n \log n + O(n),$$

and thus we have, combining (4), (5) and (7),

$$E(C) \geqq \frac{3}{2} n \log n + O(n). \tag{8}$$

3. The upper bound

We first note that

$$P(C \geqq u) = P\left(\left(\max_{1 \leqq i \leqq n} T_i\right) \geqq u\right) \leqq nP(T_1 \geqq u).$$

Now, using the well-known formula $E(C) = \int_0^\infty P(C \geqq u)du$ ([2], p. 148), we see that, for any $t > 0$,

$$\begin{aligned} E(C) &= \int_0^t P(C \geqq u)du + \int_t^\infty P(C \geqq u)du \\ &\leqq t + n \int_t^\infty P(T_1 \geqq u)du. \end{aligned} \tag{9}$$

The exact distribution of T_1 is known from Section 2, and thus

$$P(T_1 \geqq u) \leqq e^{-us} E(e^{sT_1}) = e^{-us} \prod_{i=1}^{n-1} p_i e^s/(1 - q_i e^s), \tag{10}$$

where $p_i = (2i(n-i))/(n(n-1))$, and $q_i = 1 - p_i$. In (10) we must have $1 - q_1 e^s > 0$ so $e^{-s} > 1 - 2/n$. Write $\pi_n = \Pi_{i=1}^{n-1} p_i e^s/(1 - q_i e^s)$ and then substitute (10) into (9) to obtain

$$E(C) \leqq t + ns^{-1} e^{-st} \pi_n. \tag{11}$$

Setting $e^{-s} = 1 - c/n$ with $c < 2$, we have

$$\log \pi_n = -\sum_{i=1}^{n-1} \log(1 - c/(np_i)) = \sum_{i=1}^{n-1} \sum_{j=1}^{\infty} j^{-1}(c/(np_i))^j. \tag{12}$$

If we set $b_j = \Sigma_{i=1}^{n-1} (c/(np_i))^j$, we have (as in [5]), $b_1 \leqq c \log n$, and

$$\sum_{j=2}^{\infty} j^{-1} b_j < \sum_{i=1}^{n-1} \sum_{j=2}^{\infty} (c/2)^j \left(\frac{n-1}{i(n-i)}\right)^j < \sum_{i=1}^{n-1} \left(\frac{c}{2} \frac{n-1}{i(n-i)}\right)^2 (1 - c/2)^{-1},$$

since $(n-1)/(i(n-i)) \leqq 1$ for $1 \leqq i \leqq n-1$. Now

$$\sum_{i=1}^{n-1} \left(\frac{n-1}{i(n-i)}\right)^2 = \sum_{i=1}^{n-1} \left\{\frac{1}{i^2} + \frac{1}{n}\left(\frac{1}{i} + \frac{1}{n-i}\right) + \frac{1}{(n-i)^2}\right\} = O(1),$$

so that, by (12), there is a constant M for which

$$\log \pi_n \leqq c \log n + M(1 - c/2)^{-1}.$$

Since $s > 1 - e^{-s} = c/n$, (11) now becomes

$$E(C) \leqq t + c^{-1}\exp\{2\log n - n^{-1}ct + c\log n + M(1-c/2)^{-1}\}. \tag{13}$$

On setting $c = 2(1-(\log n)^{-1/2})$ and $t = \frac{3}{2}n\log n + A\,n(\log n)^{1/2}$, the exponent in (13) becomes

$$\log n - (2A - 1 - M)(\log n)^{1/2} + 2A = \log n + 2A,$$

if we choose $A = (M+1)/2$. Then (13) becomes

$$\begin{aligned} E(C) &\leqq \frac{3}{2}n\log n + An(\log n)^{1/2} + c^{-1}n\exp(2A) \\ &= \frac{3}{2}n\log n + O(n(\log n)^{1/2}), \end{aligned} \tag{14}$$

which completes the proof of the theorem.

References

[1] Bumby, R. T. and Spencer, J. Unpublished paper cited in [2].

[2] Feller, W. (1966) *An Introduction to Probability Theory and its Applications.* Wiley, New York.

[3] Hajnal, A., Milner, E. C. and Szemeredi, E. (1972) A cure for the telephone disease. *Canad. Math. Bull.* **15**, 447–450.

[4] Moon, J. W. (1972) Random exchanges of information. *Nieuw Arch. Wisk.* **20**, 246–249.

[5] Tijdeman, R. (1971) On a telephone problem. *Nieuw Arch. Wisk.* **19**, 188–192.