CGTM 67
W.F. Miller
Lance J. Hoffman
April 1969

A Method of Extracting Record-Specific Information from "Statistical" Data Banks

The purpose of this memorandum is to display an algorithm which can be employed in conjunction with other information to obtain private information from certain classes of data banks.

Let us assume a data bank search algorithm with the following properties is provided initially:

1. The search algorithm will return to the inquirer the number of instances (people) with given properties P_1, P_2 ..., P_n but not the names or other identifying information about the number of people included in this count. That is, the algorithm permits one to get aggregate data in the form of the count of the number of instances of a certain type, but it does not return anything other than the count.
2. The search algorithm will permit requests not only for a count of the number of cases of a given property, but also for a count of the number of cases with a conjunction of properties. For example, one might ask for the number of people with property P_1 (age greater than 30) and property P_2 (female) and property P_3 (not living in the city of Palo Alto).

The following algorithm determines whether a person (called Mr. X) has property P_0 given that we a priori know a large number of his other properties. Denote the number of people with properties P_1, P_2,..., P_m in common by $\#(P_1 \,\&\, P_2 \,\&\, \ldots \,\&\, P_m)$. Use the search algorithm to determine

$$\#(P_1 \,\&\, P_2 \,\&\, \ldots\, P_i), \quad \text{for } i=1,2,3,\ldots,N_o,$$

where N_o is the smallest integer such that

$$\#(P_1 \,\&\, P_2 \,\&\, \ldots\, P_{N_o}) = 1$$

Then

if $\#(P_1 \ \& \ P_2 \ \& \ \ldots \ P_{N_o} \ \& \ P_0) = 1$, Mr. X has property P_0.

Otherwise Mr. X does not have property P_0.

The scheme will fail if we do not know enough about Mr. X to identify him through his properties $P_1, P_2, \ldots, P_N$, i.e. if $\#(P_1 \ \& \ P_2 \ \& \ \ldots \ \& \ P_N) > 1$.

There is a variation of this scheme. Suppose we know that Mr. X is included in the count of people with properties $P_1, P_2, P_3, \ldots, P_N$ in common. If the count of people with properties $P_1, P_2, \ldots, P_N, P_o$ in common is the same, i.e., if

$$\#(P_1 \ \& \ P_2 \ \& \ P_3 \ \& \ \ldots \ \& \ P_N) = \#(P_1 \ \& \ P_2 \ \& \ P_3 \ \& \ \ldots \ \& \ P_N \ \& \ P_0)$$

then we know that Mr. X has property P_0. For this variation of the scheme to work in practice, the count will surely have to be small in order that one can expect all members in the count to also have P_0.

If

$$\#(P_1 \ \& \ P_2 \ \ldots \ \& \ P_N) \neq \#(P_1 \ \& \ P_2 \ \ldots \ \& \ P_N \ \& \ P_0)$$

we cannot determine whether Mr. X has property P_0 unless we have the first case where

$$\#(P_1 \ \& \ P_2 \ \ldots \ \& \ P_N) = 1.$$

As an example of the use of this algorithm, suppose that we wish to determine whether Mr. X has been married twice, and we know that he has properties $P_1, P_2, \ldots, P_7$ where

P_1 = age 39
P_2 = education level is B.S.
P_3 = male
P_4 = has 4 children
P_5 = lives in Palo Alto
P_6 = profession is lawyer
P_7 = salary exceeds $25,000.00 per year
P_8 = has been married twice
.
.
.

Then if $\#(P_1 \& P_2 \& \ldots \& P_7) = 1$, Mr. X has been married twice if and only if $\#(P_1 \& P_2 \& \ldots \& P_7 \& P_8) = 1$. We can thus obtain information on a specific individual even though the search algorithm provided only returns counts of instances of conjunctions of properties.