A THEORY FOR RECORD LINKAGE*

IVAN P. FELLEGI AND ALAN B. SUNTER
Dominion Bureau of Statistics

A mathematical model is developed to provide a theoretical framework for a computer-oriented solution to the problem of recognizing those records in two files which represent identical persons, objects or events (said to be *matched*).

A comparison is to be made between the recorded characteristics and values in two records (one from each file) and a decision made as to whether or not the members of the comparison-pair represent the same person or event, or whether there is insufficient evidence to justify either of these decisions at stipulated levels of error. These three decisions are referred to as *link* (A_1), a *non-link* (A_3), and a *possible link* (A_2). The first two decisions are called positive dispositions.

The two types of error are defined as the error of the decision A_1 when the members of the comparison pair are in fact unmatched, and the error of the decision A_3 when the members of the comparison pair are, in fact matched. The probabilities of these errors are defined as

$$\mu = \sum_{\gamma \in \Gamma} u(\gamma) P(A_1 \mid \gamma)$$

and

$$\lambda = \sum_{\gamma \in \Gamma} m(\gamma) P(A_3 \mid \gamma)$$

respectively where $u(\gamma)$, $m(\gamma)$ are the probabilities of realizing γ (a comparison vector whose components are the coded agreements and disagreements on each characteristic) for unmatched and matched record pairs respectively. The summation is over the whole comparison space Γ of possible realizations.

A *linkage rule* assigns probabilities $P(A_1|\gamma)$, and $P(A_2|\gamma)$, and $P(A_3|\gamma)$ to each possible realization of $\gamma \in \Gamma$. An optimal linkage rule $L(\mu, \lambda, \Gamma)$ is defined for each value of (μ, λ) as the rule that minimizes $P(A_2)$ at those error levels. In other words, for fixed levels of error, the rule minimizes the probability of failing to make positive dispositions.

A theorem describing the construction and properties of the optimal linkage rule and two corollaries to the theorem which make it a practical working tool are given.

1. INTRODUCTION

THE necessity for comparing the records contained in a file L_A with those in a file L_B in an effort to determine which pairs of records relate to the same population unit is one which arises in many contexts, most of which can be categorized as either (a) the construction or maintenance of a master file for a population, or (b) merging two files in order to extend the amount of information available for population units represented in both files.

The expansion of interest in the problem in the last few years is explained by three main factors:

1) the creation, often as a by-product of administrative programmes, of large files which require maintenance over long periods of time and which often contain important statistical information whose value could be increased by linkage of individual records in different files;

*Reprinted with permission from the Journal of the American Statistical Association, American Statistical Association, December 1969, Vol. 64, No. 328, pp. 1183-1210.

2) increased awareness in many countries of the potential of record linkage for medical and genetic research;
3) advances in electronic data processing equipment and techniques which make it appear technically and economically feasible to carry out the huge amount of operational work in comparing records between even medium-sized files.

A number of computer-oriented record linkage operations have already been reported in the literature ([4], [5], [6], [7], [8], [11], [12], [13]) as well as at least two attempts to develop a theory for record linkage ([1], [3]). The present paper is, the authors hope, an improved version of their own earlier papers on the subject ([2], [9], [10]). The theory, developed along the lines of classical hypothesis testing, leads to a linkage rule which is quite similar to the intuitively appealing approach of Newcombe ([4], [5], [6]).

The approach of the present paper is to create a mathematical model within the framework of which a theory is developed to provide guidance for the handling of the linkage problem. Some simplifying assumptions are introduced and some practical problems are examined.

2. THEORY

There are two populations A and B whose elements will be denoted by a and b respectively. We assume that some elements are common to A and B. Consequently the set of ordered pairs

$$A \times B = \{(a, b); a \epsilon A, b \epsilon B)$$

is the union of two disjoint sets

$$M = \{(a, b); a = b, a \epsilon A, b \epsilon B\} \tag{1}$$

and

$$U = \{(a, b); a \neq b, a \epsilon A, b \epsilon B\} \tag{2}$$

which we call the *matched* and *unmatched* sets respectively.

Each unit in the population has a number of characteristics associated with it (e.g. name, age, sex, marital status, address at different points in time, place and date of birth, etc.). We assume now that there are two record generating processes, one for each of the two populations. The result of a record generating process is a record for each member of the population containing some selected characteristics (e.g. age at a certain date, address at a certain date, etc.). The record generating process also introduces some errors and some incompleteness into the resulting records (e.g. errors of reporting or failure to report, errors of coding, transcribing, keypunching, etc.). As a result two unmatched members of A and B may give rise to identical records (either due to errors or due to the fact that an insufficient number of characteristics are included in the record) and, conversely, two matched (identical) members of A and B may give rise to different records. We denote the records corresponding to members of A and B by $\alpha(a)$ and $\beta(b)$ respectively.

We also assume that simple random samples, denoted by A_s and B_s respectively, are selected from each of A and B. We do not, however, exclude the

possibility that $A_s = A$ and $B_s = B$. The two given files, L_A and L_B, are considered to be the result of the application of the record generating process to A_s and B_s respectively. For simplicity of notation we will drop the subscript s.

The first step in attempting to link the records of the two files (i.e. identifying the records which correspond to matched members of A and B) is the comparison of records. The result of comparing two records, is a set of codes encoding such statements as "name is the same," "name is the same and it is Brown," "name disagrees," "name missing on one record," "agreement on city part of address, but not on street," etc. Formally we define the *comparison vector* as a vector function of the records $\alpha(a)$, $\beta(b)$:

$$\gamma[\alpha(a), \beta(b)] = \{\gamma^1[\alpha(a), \beta(b)], \cdots, \gamma^K[\alpha(a), \beta(b)]\} \tag{3}$$

It is seen that γ is a function on $A \times B$. We shall write $\gamma(a, b)$ or $\gamma(\alpha, \beta)$ or simply γ as it serves our purpose. The set of all possible realizations of γ is called the *comparison space* and denoted by Γ.

In the course of the linkage operation we observe $\gamma(a, b)$ and want to decide either that (a, b) is a matched pair $(a, b) \in M$ (call this decision, denoted by A_1, a *positive link*) or that (a, b) is an unmatched pair $(a, b) \in U$ (call this decision, denoted by A_3, a *positive non-link*). There will be however some cases in which we shall find ourselves unable to make either of these decisions at specified levels of error (as defined below) so that we allow a third decision, denoted A_2, a *possible link*.

A *linkage rule* L can now be defined as a mapping from Γ, the comparison space, onto a set of random decision functions $D = \{d(\gamma)\}$ where

$$d(\gamma) = \{P(A_1 \mid \gamma), P(A_2 \mid \gamma), P(A_3 \mid \gamma)\}; \quad \gamma \in \Gamma \tag{4}$$

and

$$\sum_{i=1}^{3} P(A_i \mid \gamma) = 1. \tag{5}$$

In other words, corresponding to each observed value of γ, the linkage rule assigns the probabilities for taking each of the three possible actions. For some or even all of the possible values of γ the decision function may be a degenerate random variable, i.e. it may assign one of the actions with probability equal to 1.

We have to consider the levels of error associated with a linkage rule. We assume, for the time being, that a pair of records $[\alpha(a), \beta(b)]$ is selected for comparison according to some probability process from $L_A \times L_B$ (this is equivalent to selecting a pair of elements (a, b) at random from $A \times B$, due to the construction of L_A and L_B). The resulting comparison vector $\gamma[\alpha(a), \beta(b)]$ is a random variable. We denote the conditional probability of γ, given that $(a, b) \in M$ by $m(\gamma)$. Thus

$$\begin{aligned} m(\gamma) &= P\{\gamma[\alpha(a), \beta(b)] \mid (a, b) \in M\} \\ &= \sum_{(a,b) \in M} P\{\gamma[\alpha(a), \beta(b)]\} \cdot P[(a, b) \mid M]. \end{aligned} \tag{6}$$

Similarly we denote the conditional probability of γ, given that $(a, b) \in U$ by $u(\gamma)$. Thus

$$u(\gamma) = P\{\gamma[\alpha(a), \beta(b)] \mid (a, b) \in U\} = \sum_{(a,b)\in U} P\{\gamma[\alpha(a), \beta(b)]\} \cdot P[(a, b) \mid U]. \tag{7}$$

There are two types of error associated with a linkage rule. The first occurs when an unmatched comparison is linked and has the probability

$$P(A_1 \mid U) = \sum_{\gamma \in \Gamma} u(\gamma) \cdot P(A_1 \mid \gamma). \tag{8}$$

The second occurs when a matched comparison is non-linked and has the probability

$$P(A_3 \mid M) = \sum_{\gamma \in \Gamma} m(\gamma) \cdot P(A_3 \mid \gamma). \tag{9}$$

A linkage rule on the space Γ will be said to be a linkage rule at the levels μ, λ ($0 < \mu < 1$ and $0 < \lambda < 1$) and denoted by $L(\mu, \lambda, \Gamma)$ if

$$P(A_1 \mid U) = \mu \tag{10}$$

and

$$P(A_3 \mid M) = \lambda. \tag{11}$$

Among the class of linkage rules on Γ which satisfy (10) and (11) the linkage rule $L(\mu, \lambda, \Gamma)$ will be said to be the *optimal linkage rule* if the relation

$$P(A_2 \mid L) \leqq P(A_2 \mid L') \tag{12}$$

holds for every $L'(\mu, \lambda, \Gamma)$ in the class.

In explanation of our definition we note that the optimal linkage rule maximizes the probabilities of positive dispositions of comparisons (i.e. decisions A_1 and A_3) subject to the fixed levels of error in (10) and (11) or, put differently, it minimizes the probability of failing to make a positive disposition. This seems a reasonable approach since in applications the decision A_2 will require expensive manual linkage operations; alternatively, if the probability of A_2 is not small, the linkage process is of doubtful utility.

It is not difficult to see that for certain combinations of μ and λ the class of linkage rules satisfying (10) and (11) is empty. We admit only those combinations of μ and λ for which it is possible to satisfy equations (10) and (11) simultaneously with some set D of decision functions as defined by (4) and (5). For a more detailed discussion of admissibility see Appendix 1. At this point it is sufficient to note that a pair of values (μ, λ) will be inadmissible only if one or both of the members are too large, and that in this case we would always be happy to reduce the error levels.

2.1. *A fundamental theorem*

We first define a linkage rule L_0 on Γ. We start by defining a unique ordering of the (finite) set of possible realizations of γ.

If any value of γ is such that both $m(\gamma)$ and $u(\gamma)$ are equal to zero, then the (unconditional) probability of realizing that value of γ is equal to zero, and

hence it need not be included in Γ. We now assign an order arbitrarily to all γ for which $m(\gamma)>0$ but $u(\gamma)=0$.

Next we order all remaining γ in such a way that the corresponding sequence of

$$m(\gamma)/u(\gamma)$$

is monotone decreasing. When the value of $m(\gamma)/u(\gamma)$ is the same for more than one γ we order these γ arbitrarily.

We index the ordered set $\{\gamma\}$ by the subscript i; $(i=1, 2, \cdots, N_\Gamma)$; and write $u_i=u(\gamma_i)$; $m_i=m(\gamma_i)$.

Let (μ, λ) be an admissible pair of error levels and choose n and n' such that

$$\sum_{i=1}^{n-1} u_i < \mu \leqq \sum_{i=1}^{n} u_i \tag{13}$$

$$\sum_{i=n'}^{N_\Gamma} m_i \geqq \lambda > \sum_{i=n'+1}^{N_\Gamma} m_i \tag{14}$$

where N_Γ is the number of points in Γ.

We assume for the present that when (13) and (14) are satisfied we have $1<n\leq n'-1<N_\Gamma$. This will ensure that the levels (μ, λ) are admissible. Let $L_0(\mu, \lambda, \Gamma)$ denote the linkage rule defined as follows: having observed a comparison vector, γ_i, take action A_1 (positive link) if $i\leqq n-1$, action A_2 when $n<i\leqq n'-1$, and action A_3 (positive non-link) when $i\geqq n'+1$. When $i=n$ or $i=n'$ then a random decision is required to achieve the error levels μ and λ exactly. Formally,

$$d(\gamma_i) = \begin{cases} (1, 0, 0) & i \leq n-1 \quad \text{(a)} \\ (P_\mu, 1-P_\mu, 0) & i = n \quad \text{(b)} \\ (0, 1, 0) & n < i \leq n'-1 \quad \text{(c)} \\ (0, 1-P_\lambda, P_\lambda) & i = n' \quad \text{(d)} \\ (0, 0, 1) & i \geq n'+1 \quad \text{(e)} \end{cases} \tag{15}$$

where P_μ and P_λ are defined as the solutions to the equations

$$u_n \cdot P_\mu = \mu - \sum_{i=1}^{n-1} u_i \tag{16}$$

$$m_{n'} \cdot P_\lambda = \lambda - \sum_{i=n'+1}^{N_\Gamma} m_i. \tag{17}$$

THEOREM[1]: Let $L_0(\mu, \lambda, \Gamma)$ be the linkage rule defined by (15). Then L is a best linkage rule on Γ at the levels (μ, λ). The proof is given in Appendix 1.

The reader will have observed that the whole theory could have been formulated, although somewhat awkwardly, in terms of the classical theory of hypothesis testing. We can test first the null hypothesis that $(a, b)\in U$ against

[1] A slightly extended version of the theorm is given in Appendix 1.

the simple alternative that $(a, b) \in M$, the action A_1 being the rejection of the null hypothesis and μ the level of significance. Similarly the action A_3 is the rejection at the significance level λ of the null hypothesis that $(a, b) \in M$ in favour of the simple alternative that $(a, b) \in U$. The linkage rule L is equivalent to the likelihood ratio test and the theorem above asserts this to be the uniformly most powerful test for either hypothesis.

We state, without proof, two corollaries to the theorem. These corollaries, although mathematically trivial, are important in practice.

Corollary 1: If

$$\mu = \sum_{i=1}^{n} u_i, \quad \lambda = \sum_{i=n}^{N\Gamma} m_i, \quad n < n',$$

the $L_0(u, \lambda, \Gamma)$, the best linkage rule at the levels (μ, λ) becomes

$$d(\gamma_i) = \begin{cases} (1,0,0) & \text{if} \quad 1 \leq i \leq n \\ (0,1,0) & \text{if} \quad n < i < n' \\ (0,0,1) & \text{if} \quad n' \leq i \leq N_\Gamma. \end{cases} \tag{18}$$

If we define

$$T_\mu = \frac{m(\gamma_n)}{u(\gamma_n)}$$

$$T_\lambda = \frac{m(\gamma_{n'})}{u(\gamma_{n'})}$$

then the linkage rule (18) can be written equivalently[2] as

$$d(\gamma) = \begin{cases} (1,0,0) & \text{if} \quad T_\mu \leq m(\gamma)/u(\gamma) \\ (0,1,0) & \text{if} \quad T_\lambda < m(\gamma)/u(\gamma) < T_u \\ (0,0,1) & \text{if} \quad m(\gamma)/u(\gamma) \leq T_\lambda. \end{cases} \tag{19}$$

Corollary 2: Let T_μ and T_λ be any two positive numbers such that

$$T_\mu > T_\lambda.$$

Then there exists an admissible pair of error levels (μ, λ) corresponding to T_μ and T_λ such that the linkage rule (19) is best at these levels. The levels (μ, λ) are given by

$$\mu = \sum_{\gamma \in \Gamma_\mu} u(\gamma) \tag{20}$$

$$\lambda = \sum_{\gamma \in \Gamma_\lambda} m(\gamma) \tag{21}$$

where

$$\Gamma_\mu = \{\gamma : T_\mu \leq m(\gamma)/u(\gamma)\} \tag{22}$$

$$\Gamma_\lambda \{\gamma : m(\gamma)/u(\gamma) \leq T_\lambda\} \tag{23}$$

---

[2] We are grateful to the referee for pointing out that (19) and (18) are exactly equivalent only if $m_n/u_n < m_{n+1}/u_{n+1}$ and $m_{n'-1}/u_{n'-1} < m_{n'}/u_{n'}$.

In many applications we may be willing to tolerate error levels sufficiently high to preclude the action A_2. In this case we choose n and n' or, alternatively, T_μ and T_λ so that the middle set of γ in (18) or (19) is empty. In other words every (a, b) is allocated either to M or to U. The theory for the allocation of observations to one of two mutually exclusive populations may thus be regarded as a special case of the theory given in this paper.

3. APPLICATIONS

3.1. *Some Practical Problems*

In attempting to implement the theory developed in the previous section several practical problems need to be solved. They are outlined briefly below and taken up in more detail in subsequent sections.

a) The large number of possible values of $m(\gamma)$ and $u(\gamma)$. Clearly the number of distinct realizations of γ may be so large as to make the computation and storage of the corresponding values of $m(\gamma)$ and $u(\gamma)$ impractical. The amount of computation and storage can be substantially reduced on the basis of some simplifying assumptions.

b) Methods to calculate the quantities $m(\gamma)$ and $u(\gamma)$. Two methods are proposed.

c) Blocking the files. Implicit in the development of the theory is the assumption that if two files are linked then all possible comparisons of all the records of both files will be attempted. It is clear that even for medium sized files the number of comparisons under this assumption would be very large, (e.g. 10^5 records in each file would imply 10^{10} comparisons). In practice the files have to be "blocked" in some fashion and comparisons made only within corresponding blocks. The impact of such blocking on the error levels will be examined.

d) Calculations of threshold values. It should be clear from Corollary 2 that we do not have to order explicitly the values of γ in order to apply the main theorem since for any particular γ the appropriate decision (A_1, A_2 or A_3) can be made by comparing $m(\gamma)/u(\gamma)$ with the threshold values T_μ and T_λ. We shall outline a method of establishing these threshold values corresponding to the required error levels μ and λ.

e) Choice of the comparison space. The main theorem provides an optimal linkage rule for a given comparison space. Some guidance will be provided on the choice of the comparison space.

3.2. *Some simplifying assumptions*

In practice the set of distinct (vector) values of γ may be so large that the estimation of the corresponding probabilities $m(\gamma)$ and $u(\gamma)$ becomes completely impracticable. In order to make use of the theorem it will be necessary to make some simplifying assumptions about the distribution of γ.

We assume that the components of γ can be re-ordered and grouped in such a way that

$$\gamma = (\gamma^1, \gamma^2, \cdots, \gamma^K)$$

and that the (vector) components are mutually statistically independent with

respect to each of the conditional distributions. Thus

$$m(\gamma) = m_1(\gamma^1) \cdot m_2(\gamma^2) \cdot \cdots m_k(\gamma^K) \tag{24}$$

$$u(\gamma) = u_1(\gamma^1) \cdot u_2(\gamma^2) \cdot \cdots u_k(\gamma^K) \tag{25}$$

where $m(\gamma)$ and $u(\gamma)$ are defined by (4) and (5) respectively and

$$m_i(\gamma^i) = P(\gamma^i \mid (a, b) \in M)$$

$$u_i(\gamma^i) = P(\gamma^i \mid (a, b) \in U).$$

For simplicity of notation we shall write $m(\gamma^i)$ and $u(\gamma^i)$ instead of the technically more precise $m_i(\gamma^i)$ and $u_i(\gamma^i)$. As an example, in a comparison of records relating to persons γ^1 might include all comparison components that relate to surnames, γ^2 all comparison components that relate to addresses. The components γ^1 and γ^2 are themselves vectors; the subcomponents of γ^2 for example might represent the coded results of comparing the different components of the address (city name, street name, house number, etc.). If two records are matched (i.e. when in fact they represent the same person or event), then a disagreement configuration could occur due to errors. Our assumption says that errors in names, for example, are independent of errors in addresses. If two records are unmatched (i.e. when in fact they represent different persons or events) then our assumption says that an accidental agreement on name, for example, is independent of an accidental agreement on address. In other words what we do assume is that $\gamma^1, \gamma^2, \cdots, \gamma^K$ are conditionally independently distributed. We emphasize that we do *not* assume anything about the unconditional distribution of γ.

It is clear that any monotone increasing function of $m(\gamma)/u(\gamma)$ could serve equally well as a test statistic for the purpose of our linkage rule. In particular it will be advantageous to use the logarithm of this ratio and define

$$w^k(\gamma^k) = \log m(\gamma^k) - \log u(\gamma^k). \tag{26}$$

We can then write

$$w(\gamma) = w^1 + w^2 + \cdots + w^K \tag{27}$$

and use $w(\gamma)$ as our test statistic with the understanding that if $u(\gamma) = 0$ or $m(\gamma) = 0$ then $w(\gamma) = +\infty$ (or $w(\gamma) = -\infty$) in the sense that $w(\gamma)$ is greater (or smaller) than any given finite number.

Suppose that γ^k can take on n_k different *configurations*, $\gamma_1^k, \gamma_2^k, \cdots, \gamma_{n_k}^k$. We define

$$w_j^k = \log m(\gamma_j^k) - \log u(\gamma_j^k). \tag{28}$$

It is a convenience for the intuitive interpretation of the linkage process that the weights so defined are positive for those configurations for which $m(\gamma_j^k) > u(\gamma_j^k)$, negative for those configurations for which $m(\gamma_j^k) < u(\gamma_j^k)$, and that this property is preserved by the weights associated with the total configuration γ.

The number of total configurations (i.e. the number of points $\gamma \in \Gamma$) is obviously $n_1 \cdot n_2 \cdot \cdots n_K$. However, because of the additive property of the

weights defined for components it will be sufficient to determine $n_1+n_2+\cdots+n_K$ weights. We can then always determine the weight associated with any γ by employing this additivity.

3.3. *The Calculation of Weights*

An assumption made at the outset of this paper was that the files L_A and L_B represent samples A_s and B_s of the populations A and B. This assumption is often necessary in some applications when one wishes to use a set of values of $m(\gamma^k)$ and $u(\gamma^k)$, computed for some large populations A and B while the actually observed files L_A and L_B correspond to some subpopulations A_s and B_s. For example, in comparing a set of incoming records against a master file in order to update the file one may want to consider the master file and the incoming set of records as corresponding to samples A_s and B_s of some conceptual populations A and B. One might compute the weights for the full comparison space Γ corresponding to A and B and apply these weights repeatedly on different update runs; otherwise one would have to recompute the weights on each occasion.

Of course it seldom occurs in practice that the subpopulations represented by the files L_A and L_B are actually drawn at random from any real populations A and B. However it is clear that all the theory presented in this paper will still hold if the assumption is relaxed to the assumption that the condition of entry of the subpopulation into the files is uncorrelated with the distribution in the populations of the characteristics used for comparisons. This second assumption obviously holds if the first does, although the converse is not necessarily true.

In this paper we propose two methods for calculating weights. In the first of these we assume that prior information is available on the distribution in the populations A and B of the characteristics used in comparison as well as on the probabilities of different types of error introduced into the files by the record generating processes. The second method utilizes the information in the files L_A and L_B themselves to estimate the probabilities $m(\gamma^k)$ and $u(\gamma^k)$. The validity of these estimates is strongly predicated on the independence assumption of the previous section. Specifically it requires that the formal expression for that independence should hold almost exactly in the subpopulation $L_A \times L_B$, which, in turn, requires that the files L_A and L_B should be large and should satisfy at least the weaker of the assumptions of the previous paragraph.

Another procedure, proposed by Tepping ([11], [13]), is to draw a sample from $L_A \times L_B$, identify somehow (with negligible error) the matched and unmatched comparisons in this sample, and thus estimate $m(\gamma)$ and $u(\gamma)$ directly. The procedure seems to have some difficulties associated with it. If and when the identification of matched and unmatched records can in fact be carried out with reasonable accuracy and with reasonable economy (even if only at least occasionally) then it might provide a useful check or corroboration of the reasonableness of assumptions underlying the calculation of weights.

Finally, the weights $w(\gamma)$ or alternatively the probabilities $m(\gamma)$ and $u(\gamma)$, derived on one occasion for the linkage $L_A \times L_B$ can continue to be used on a

subsequent occasion for the linkage, say $L_A' \times L_B'$, provided A_s and B_s can be regarded as samples from the same populations as A_s and B_s and provided the record generating processes are unaltered.

3.3.1. *Method I*

Suppose that one component of the records associated with each of the two populations A and B is the surname. The comparison of surnames on two records will result in a component of the comparison vector. This component may be a simple comparison component such as "name agrees" or "name disagrees" or "name missing on one or both records" (in this case γ^k is a scalar); or it may be a more complicated vector component such as for example "records agree on Soundex code, the Soundex code is B650; the first 5 characters of the name agree; the second 5 characters of the name agree; the surname is BROWNING."

In either of the two files the surname may be reported in error. Assume that we could list all error-free realizations of all surnames in the two populations and also the number of individuals in the respective populations corresponding to each of these surnames. Let the respective frequencies in A and B be

$$f_{A_1}, f_{A_2}, \cdots, f_{A_m}; \qquad \sum_{j=1}^{m} f_{A_j} = N_A$$

and

$$f_{B_1}, f_{B_2}, \cdots, f_{B_m}; \qquad \sum_{j=1}^{m} f_{Bj} = N_B.$$

Let the corresponding frequencies in $A \cap B$ be

$$f_1, f_2, \cdots, f_m; \qquad \sum_{j} f_j = N_{AB}.$$

The following additional notation is needed:

e_A or e_B the respective probabilities of a name being misreported in L_A or L_B (we assume that the probability of misreporting is independent of the particular name);

e_{A0} or e_{B0} the respective probabilities of a name not being reported in L_A or L_B (we assume that the probability of name not being reported is independent of the particular name);

e_T the probability the name of a person is differently (though correctly) reported in the two files (this might arise, for example, if L_A and L_B were generated at different times and the person changed his name).

Finally we assume that e_A and e_B are sufficiently small that the probability of an agreement on two identical, though erroneous, entries is negligible and that the probabilities of misreporting, not reporting and change are independent of one another.

We shall first give a few rules for the calculation of m and u corresponding

to the following configurations of γ: name agrees and it is the jth listed name, name disagrees; name missing on either record.

m (name agrees and is the jth listed name)

$$= \frac{f_j}{N_{AB}}(1-e_A)(1-e_B)(1-e_T)(1-e_{A0})(1-e_{B0})$$
$$\doteq \frac{f_j}{N_{AB}}(1-e_A-e_B-e_T-e_{A0}-e_{B0}) \tag{29}$$

m (name disagrees)

$$= [1-(1-e_A)(1-e_B)(1-e_T)](1-e_{A0})(1-e_{B0})$$
$$\doteq e_A+e_B+e_T \tag{30}$$

m (name missing on either file)

$$= 1-(1-e_{A0})(1-e_{B0}) \doteq e_{A0}+e_{B0} \tag{31}$$

u (name agrees and is the jth listed name)

$$= \frac{f_{Aj}}{N_A}\frac{f_{Bj}}{N_B}(1-e_A)(1-e_T)(1-e_{A0})(1-e_{B0})$$
$$\doteq \frac{f_{Aj}}{N_A}\frac{f_{Bj}}{N_B}(1-e_A-e_B-e_T-e_{A0}-e_{B0}) \tag{32}$$

u (name disagrees)

$$= \left[1-(1-e_A)(1-e_B)(1-e_T)\sum_j \frac{f_{Aj}}{N_A}\frac{f_{Bj}}{N_B}\right](1-e_{A0})(1-e_{B0})$$
$$\doteq \left[1-(1-e_A-e_B-e_T)\sum_j \frac{f_{Aj}}{N_A}\frac{f_{Bj}}{N_B}\right](1-e_{A0}-e_{B0}) \tag{33}$$

u (name missing on either file)

$$= 1-(1-e_{A0})(1-e_{B0}) = e_{A0}+e_{B0}. \tag{34}$$

The proportions $f_{Aj}/N_A, f_{Bj}/N_B, f_j/N$ may be taken, in many applications, to be the same. This would be the case, for example, if two large files can be assumed to be drawn from the same population. These frequencies may be estimated from the files themselves.

A second remark relates to the interpretation of weights. It will be recalled that according to (28) the contribution to the overall weight of the name component is equal to $\log(m/u)$ and that comparisons with a weight higher than a specified number will be considered linked, while those whose weight is below a specified number will be considered unlinked. It is clear from (29–34) that an agreement on name will produce a positive weight and in fact the rarer the name, the larger the weight; a disagreement on name will produce a negative weight which decreases with the errors e_A, e_B, e_T; if the name is missing on either record, the weight will be zero. These results seem intuitively appealing.

We should emphasize that it is not necessary to list all possible names for the validity of formulae (29) to (34). We might only list the more common names separately, grouping all the remaining names. In the case of groupings the appropriate formulae in (29) to (34) have to be summed over the corresponding values of the subscript j. The problem of how to group configurations is taken up in a later section.

Finally we should mention that formulae (29) to (34) relate to reasonably simple realizations of γ, such as a list of names, or list of ages, or lists of other possible identifiers. In more complex cases one may be able to make use of these results, with appropriate modifications, in conjunction with the elementary rules of probability calculus. Alternatively one may have recourse to the method given below.

3.3.2. *Method II*

The formulae presented in Appendix 2 can be used, under certain circumstances, to estimate the quantities $m(\gamma^k)$, $u(\gamma^k)$ and N, the number of matched records, simply by substituting into these formulae certain frequencies which can be directly (and automatically) counted by comparing the two files. Mathematically, the only condition for the validity of these formulae is that γ should have at least three components which are independent with respect to the probability measures m and u in the sense of (24) and (25). It should be kept in mind, however, that for agreement configurations $m(\gamma^k)$ is typically very close to one, $u(\gamma^k)$ is very close to zero, and conversely for diagreement configurations. Therefore the estimates of $u(\gamma^k)$ and $m(\gamma^k)$ can be subject to substantial sampling variability unless the two files represent censuses or large random samples of the populations A and B.

The detailed formulae and their proofs are included in the Appendix. At this point only an indication of the methods will be given. For simplicity we present the method in terms of three components. If, in fact, there are more than three components they can be grouped until there are only three left. Clearly this can be done without violating (24) and (25).

For each component vector of γ designate the set of configurations to be considered as "agreements" and denote this set (of vectors) for the hth component by S_h. The designation of specific configurations as "agreements" may be arbitrary but subject to some numerical considerations to be outlined in the Appendix.

The following notation refers to the frequencies of various configurations of γ. Since they are not conditional frequencies, they can be obtained as direct counts by comparing the files L_A and L_B:

M_h: the proportion of "agreement" in all components except the hth; any configuration in the kth component;

U_h: the proportion of "agreement" in the hth component; any configuration in the others;

M: the proportion of "agreement" in all components.

Denote also the respective conditional probabilities of "agreements" by

$$m_h = \sum_{\gamma \in S_h} m(\gamma) \tag{35}$$

$$u_h = \sum_{\gamma \in S_h} u(\gamma). \tag{36}$$

It follows from the assumptions (24) and (25) that the expected values of M_h, U_h, and M with respect to the sampling procedure (if any) and the record generating process through which the files L_A and L_B arose from the populations A and B can be expressed simply in terms of m_h and u_h as follows.

$$N_A N_B E(M_h) = E(N) \prod_{\substack{j=1 \\ j \neq h}}^{3} m_j + [N_A N_B - E(N)] \prod_{\substack{j=1 \\ j \neq h}}^{3} u_\delta; \qquad h = 1, 2, 3 \tag{37}$$

$$N_A N_B E(U_h) = E(N) m_h + [N_A N_B - E(N)] u_h \tag{38}$$

$$N_A N_B E(M) = E(N) \prod_{j=1}^{3} m_j + [N_A N_B - E(N)] \prod_{j=1}^{3} u_j \tag{39}$$

where N_A and N_B are the known number of records in the files L_A and L_B and N is the unknown number of matched records.

Dropping the expected values we obtain seven equations for the estimation of the seven unknown quantities N, m_h, $u_h (h = 1, 2, 3)$. The solution of these equations is given in Appendix 2.

Having solved for m_h, u_h and N the quantities $m(\gamma^k)$ and $u(\gamma^k)$ are easily computed by substituting some additional directly observable frequencies into some other equations, also presented in Appendix 2. The frequency counts required for all the calculations can be obtained at the price of three sorts of the two files.

It is our duty to warn the reader again that although these equations provide statistically consistent estimates, the sampling variability of the estimates may be considerable if the number of records involved ($N_A N_B$) is not sufficiently large. One might get an impression of the sampling variabilities through the method of random replication, i.e., by splitting both of the files at random into at least two parts and by performing the estimation separately for each. Alternatively, one can at least get an impression of the sampling variabilities of M_h, U_h and M by assuming that they are estimated from a random sample of size $N_A N_B$.

Another word of caution may be in order. The estimates are computed on the basis of the independence assumptions of (24) and (25). In the case of departures from independence the estimates, *as estimates of the probabilities* $m(\gamma^k)$ *and* $u(\gamma^k)$, may be seriously affected and the resulting weights $m(\gamma^k)/u(\gamma^k)$ would lose their probabilistic interpretations. What is important, of course, is their effect on the resulting linkage operation. We believe that if sufficient identifying information is available in the two files to carry out the linkage operation in the first place, then the operation is quite robust against departures from independence. One can get an impression of the extent of the departures from independence by carrying out the calculations of Appendix 2 on the basis of alternative designations of the "agreement" configurations.

3.4. *Restriction of Explicit Comparisons to a Subspace*

In practice of course we do not select comparisons at random from $L_A \times L_B$. But then in practice we are not concerned with the *probability* of the event $(A_1|\, U)$ or the event $(A_3|\, M)$ for any particular comparison but rather with the *proportion* of occurrences of these two events in the long run. Clearly if our linkage procedure is to examine *every* comparison $(\alpha, \beta) \in L_A \times L_B$ then we could formally treat any particular comparison as if it had been drawn at random from $L_A \times L_B$. The only change in our theory in this case would be the replacement of *probabilities* with *proportions*. In particular the probabilities of error μ and λ would then have to be interpreted as proportions of errors. With this understanding we can continue to use the notation and concepts of probability calculus in this paper even though often we shall think of probabilities as proportions.

We have now made explicit a second point which needs to be examined. We would seldom be prepared to examine every $(\alpha, \beta) \in L_A \times L_B$ since it is clear that even for medium sized files (say 10^5 record each) the number of comparisons (10^{10}) would outstrip the economic capacity of even the largest and fastest computers.

Thus the number of comparisons we will examine explicitly will be restricted to a subspace, say Γ^*, of Γ. This might be achieved for example by partitioning or "blocking" the two files into Soundex-coded Surname "blocks" and making explicit comparisons only between records in corresponding blocks. The subspace Γ^* is then the set of γ for which the Soundex Surname component has the agreement status. All other γ are implicit positive non-links (the comparisons in $\Gamma - \Gamma^*$ will not even be actually compared hence they may not be either positive or possible links). We consider the effect that this procedure has on the error levels established for the all-comparison procedure.

Let Γ_μ and Γ_λ be established (as in Corollary 2) for the all-comparison procedure so as to satisfy

$$\Gamma_\mu = \{\gamma : T_\mu \leq m(\gamma)/u(\gamma)\}$$
$$\Gamma_\lambda = \{\gamma : m(\gamma)/u(\gamma) \leq T_\lambda\}$$

where

$$\mu = \sum_{\gamma \in \Gamma_\mu} u(\gamma)$$
$$\lambda = \sum_{\gamma \in \Gamma_\lambda} m(\gamma).$$

If we now regard all $\gamma \in (\Gamma - \Gamma^*)$ as implicit positive non-links we must adjust our error levels to

$$\mu^* = \mu - \sum_{\Gamma_\mu \cap \bar{\Gamma}^*} u(\gamma) \tag{40}$$

$$\lambda^* = \lambda + \sum_{\bar{\Gamma}_\lambda \cap \bar{\Gamma}^*} m(\gamma) \tag{41}$$

where $\bar{\Gamma}_\lambda$ and $\bar{\Gamma}^*$ denote complements taken with respect to Γ (i.e. $\Gamma - \Gamma_\lambda$ and $\Gamma - \Gamma^*$, respectively).

The first of these expressions indicates that the level of μ is reduced by the sum of the u-probabilities of those comparisons which would have been links under the all-comparison procedure but are implicit non-links under the blocking procedure. The second expression indicates that the actual level of λ is increased by the sum of the m-probabilities of the comparisons that would be links or possible links under the all-comparison procedure but are implicit non-links under the blocking procedure.

The probabilities of a failure to make a positive disposition under the blocking procedure are given by

$$P^*(A_2 \mid M) = \sum_{\gamma \in \bar{\Gamma}_\mu \cap \bar{\Gamma}_\lambda} m(\gamma) - \sum_{\gamma \in \bar{\Gamma}_\mu \cap \bar{\Gamma}_\lambda \cap \bar{\Gamma}_*} m(\gamma) \tag{42}$$

$$P^*(A_2 \mid U) = \sum_{\gamma \in \bar{\Gamma}_\mu \cap \bar{\Gamma}_\lambda} u(\gamma) - \sum_{\gamma \in \bar{\Gamma}_\mu \cap \bar{\Gamma}_\lambda \cap \bar{\Gamma}_*} u(\gamma) \tag{43}$$

the second term on the right in each case being the reduction due to the blocking procedure.

These expressions will be found to be useful when we consider the best way of blocking a file.

3.5. *Choice of Error Levels and Choice of Subspace*

In choosing the error levels (μ, λ) we may want to be guided by the consideration of losses incurred by the different actions.

Let $G_M(A_i)$ and $G_U(A_i)$ be non-negative loss functions which give the loss associated with the disposition A_i; $(i = 1, 2, 3)$; for each type of comparison. Normally, we would set

$$G_M(A_1) = G_U(A_3) = 0$$

and we do so here. Reverting to the all-comparison procedure we set (μ, λ) so as to minimize the expected loss given by the expression

$$\begin{aligned} & P(M) \cdot E[G_M(A_i)] + P(U) \cdot E[G_U(A_i)] \\ = \; & P(M)[P(A_2 \mid M) \cdot G_M(A_2) + \lambda \cdot G_M(A_3)] \\ & + P(U)[\mu \cdot G_U(A_1) + P(A_2 \mid U) \cdot G_U(A_2)] \end{aligned} \tag{44}$$

Note that $P(A_2 \mid M)$ and $P(A_2 \mid U)$ are functions of μ and λ. We give later a practical procedure for determining the values of (μ, λ) which minimize (44).

Suppose that (μ, λ) have been set so as to minimize (44). We now consider the effects of blocking the files and introduce an additional component in the loss function which expresses the costs of comparisons, $G_{\Gamma^*}(L_A \times L_B)$, under a blocking procedure equivalent to making implicit comparisons in a subspace Γ^*. We seek that subspace Γ^* which minimizes the total expected loss,

$$\begin{aligned} & c\{P(M) \cdot E[G_M(A_i)] + P(U) \cdot E[G_U(A_i)]\} \\ & + G_{\Gamma^*}(L_A \times L_B) \\ = \; & c\{P(M)[P^*(A_2 \mid M) G_M(A_2) + \lambda^* G_M(A_3)] \\ & + P(U)[\mu^* G_U(A_1) + P^*(A_2 \mid U) G_U(A_2)]\} \\ & + G_{\Gamma^*}(L_A \times L_B) \end{aligned} \tag{45}$$

where P^* denotes probabilities under the blocking procedure given by (42) and (43) respectively and c denotes the number of comparisons in $L_A \times L_B$. Now if the processing cost of comparisons under any blocking Γ^* is simply proportional to the number of comparisons, c^*, i.e.

$$G_{\Gamma^*}(L_A \times L_B) = \alpha c^*$$

then we can minimize

$$P(M)[P^*(A_2 \mid M)G_M(A_2)\lambda^* G_M(A_3)]$$
$$+ P(U)[\mu^* G_U(A_1) + P^*(A_2 \mid U)G_U(A_2)] + \frac{\alpha c^*}{c} \cdot \tag{46}$$

The last term is the product of the cost, α, per comparison and the reduction ratio in the number of comparisons to be made explicitly.

No explicit solution of (46) seems possible under such general conditions. However, (46) can be used to compare two different choices of Γ^*. Once a choice of Γ^* has been made, the "theoretical" error levels μ, λ can be chosen, using (40) and (41), so that the actual error levels μ^*, λ^* meet the error specification. The threshold values T_μ, T_λ are then calculated from the "theoretical" error levels.

3.6. *Choice of comparison space*

Let Γ and Γ' be two comparison spaces, with conditional distributions $m(w)$, $u(w)$ and $m'(w)$, $u'(w)$ and threshold values T_μ, T_λ and T'_μ, T'_λ respectively (the threshold values being in both cases so determined that they lead to the same error levels μ, λ).

Now in a manner precisely analogous to our linkage criterion we might say that a comparison space Γ is better than a comparison space Γ' at the error levels (μ, λ) if

$$P(T_\lambda < w(\gamma) < T_\mu) < P(T'_\lambda < w'(\gamma') < T'_\mu) \tag{47}$$

where it is assumed that the comparisons are made under the optimal linkage rule in each case. The linkage criterion developed for a given Γ is independent of (μ, λ) and $P(M)$. Clearly we cannot hope for this to be the case in general with a criterion for the choice of a comparison space.

Expanding the expression (47) we have as our criterion at the level (μ, λ)

$$P(M) \cdot \sum_{T_\lambda < w < T_\mu} m(w) + P(U) \cdot \sum_{T_\lambda < w < T_\mu} u(w)$$
$$< P(M) \cdot \sum_{T_\lambda < w < T_\mu} m(w') + P(U) \cdot \sum_{T_\lambda < w' < T_\mu} u(w') \tag{48}$$

In most practical cases of course $P(M)$ is very small and the two sides of (48) are dominated by the second term. However if a "blocking" procedure has reduced the number of unmatched comparisons greatly it would be more appropriate to use $P^*(M)$ and $P^*(U)$ appropriate to the subspace Γ^* (i.e. to the set of comparisons that will be made explicitly), than to use $P(M)$ and $P(U)$ provided the same "blocking" procedure is to be used for each choice of comparison space. $P(M)$ and $P(U)$, or alternatively $P^*(M)$ and $P^*(U)$, have to be

guessed at for the application of (48). The difference between the right hand side and the left hand side of (48) is equal to the reduction of $P(A_2)$ due to the choice of the comparison space.

In practice the difference between two comparison spaces will often be the number of configurations of component vectors which are listed out in addition to the simple "agreement"—"disagreement" configurations (e.g. "agreement on name Jones," "agreement on name Smith," etc.). The formula (48) can be used to compare the loss or gain in dropping some special configurations or listing out explicitly some more.

3.7. *Calculation of threshold values*

Having specified all the relevant configurations γ_j^k and determined their associated weights w_j^k; $k=1, 2, \cdots, K$; $j=1, 2, \cdots, n_k$ it remains to set the threshold values T_μ and T_λ corresponding to given μ and λ and to estimate the number or proportion of failures to make positive dispositions of comparisons.

As shown before, the number of weights to be determined is equal to $n_1+n_2 \cdots +n_K$. The total number of different configurations is, however, $n_1 n_2 \cdots n_K$. Since the number of total configurations will, in most practical situations, be too large for their complete listing and ordering to be feasible we have resorted to sampling the configurations in order to estimate T_μ and T_λ. Since we are primarily interested in the two ends of an ordered list of total configurations we sample with relatively high probabilities for configurations which have very high or very low weights $w(\gamma)$.

The problem is made considerably easier by the independence of the component vectors γ^k. Thus if we sample independently the component configurations $\gamma_{j_1}^1, \gamma_{j_2}^2, \cdots, \gamma_{j_K}^K$ with probabilities $z_{j_1}^1, z_{j_2}^2, \cdots, z_{j_K}^K$ respectively we will have sampled the total configuration $\gamma_j=(\gamma_{j_1}^1, \gamma_{j_2}^2, \cdots, \gamma_{j_K}^K)$ with probability $z_j=z_{j_1}^1, z_{j_2}^2 \cdots z_{j_K}^K$. Hence we do not need to list all configurations of γ for sampling purposes, only all configurations of γ^k for each k.

We speed up the sampling process and increase the efficiency of the sample by ordering the configurations listed for each component by decreasing values w^k, and sampling according to the following scheme:

1) Assign selection probabilities $z_1^k, z_2^k, \cdots, z_{n_k}^k$ roughly proportional to $|w_j^k|$.
2) Choose a configuration from each component. If the configuration γ_j^k is chosen from the kth component (with probability z_j^k) choose also the configuration $\gamma_{n_k-j+1}^k$.
3) Combine the first members of the pairs chosen from each component to give one total configuration and the second members to give another.
4) Repeat the whole procedure $S/2$ times to give a with-replacement sample of S total configurations.

The sample is then ordered by decreasing values of

$$w = w_1 + w_2 + \cdots + w_K. \tag{49}$$

Let $\gamma_h (h=1, 2, \cdots, S)$ be the hth member of the ordered listing of the sample. (Note: If a configuration with the same value of w occurs twice in the sample, it is listed twice.) Then $P(w(\gamma)<w(\gamma_h)|\gamma \in M)$ is estimated by

$$\lambda_h = \sum_{h'=h}^{S} m(\gamma_{h'})/\pi(\gamma_{h'}) \tag{50}$$

where

$$\pi(\gamma_h) = \frac{S}{2} \cdot z'(\gamma_h) \tag{51}$$

and

$$z'(\gamma_h) = z^1_{h_1} z^2_{h_2} \cdots z^K_{h_K} + z^1_{n_1-h_1+1} z^2_{n_2-h_2+1} \cdots z^K_{n_K-h_K+1} \tag{52}$$

while

$$\begin{aligned} &P'(w(\gamma) < w(\gamma_h) \mid \gamma \in U) \qquad \text{is estimated by} \\ &\mu_h = \sum_{h'=1}^{h} u(\gamma_{h'})/\pi(\gamma_{h'}). \end{aligned} \tag{53}$$

The threshold values $T(\lambda_{h'})$ and $T(\mu_{h'})$, are simply the weights $w(\gamma_{h'})$ and $w(\gamma_{h'})$.

We have written a computer program which, working from a list of configurations for each vector component and associated selection probabilities, selects a sample of total configurations, orders the sample according to (49), calculates the estimates (50) and (53) and finally prints out the whole list giving for each total configuration its associated λ_h, μ_h, $T(\lambda_h)$, and $T(\mu_h)$.

We can use the same program to examine alternative blocking procedures (see Section 3.4). Thus in the ordered listing of sampled configurations we can identify those which would be implicit positive non-links under a blocking procedure which restricts explicit comparisons to a subspace Γ^*. Thus corresponding to any values of T_μ and T_λ (or μ and λ) we can obtain the second terms in each of the expressions (40), (41), (42), and (43). Alternatively if the implicit positive non-links are passed over in the summations (40) and (41) we can read off the values of the left-hand sides of those expressions. If we arrange this for alternative blocking procedures we are able to use the output of the program to make a choice of blocking procedures according to (46).

4. ACKNOWLEDGMENTS

The authors would like to express their gratitude to the Dominion Bureau of Statistics for providing opportunties for this research and in particular to Dr. S. A. Goldberg for his continued support.

The authors would also like to express their appreciation to H. B. Newcombe for his pioneering work in the field of record linkage and for his generous encouragement of an approach which, in many respects, differs from his own. The contributions of J. N. Gauthier, the systems analyst in charge of programming for our pilot project, have been essential to whatever success we have enjoyed so far and will continue to be essential in what remains for us to do.

REFERENCES

[1] Du Bois, N. S. D., "A solution to the problem of linking multivariate documents, *Journal of the American Statistical Association*, 64 (1969) 163–174.

[2] Fellegi, I. P. and Sunter, A. B., "An optimal theory of record linkage," *Proceedings of the International Symposium on Automation of Population Register Systems, Volume 1*, Jerusalem, Israel, 1967.
[3] Nathan, G., "Outcome probabilities for a record matching process with complete invariant information," *Journal of the American Statistical Association*, 62 (1967) 454–69.
[4] Newcombe, H. B. and Kennedy, J. M., "Record linkage: Making maximum use of the discriminating power of identifying information," *Communications of the A.C.M.* 5 (1962) 563.
[5] Newcombe, H. B., Kennedy, J. M., Axford, S. L., and James, A. P., "Automatic linkage of vital records," *Science* 130, (1959) 954.
[6] Newcombe, H. B. and Rhynas, P. O. W., "Family linkage of population records," Proc. U.N. / W. H. O. Seminar on Use of Vital and Health Statistics for Genetic and Radiation Studies; United Nations Sales No: 61, XVII 8, New York, 1962.
[7] Nitzberg, David M. and Sardy, Hyman, "The methodology of computer linkage of health and vital records," *Proc. Soc. Statist. Section, American Statistical Association*, Philadelphia, 1965.
[8] Phillips, Jr., William and Bahn, Anita K., "Experience with computer matching of names," *Proc. Soc. Statist. Section, American Statistical Association*, Philadelphia, 1965.
[9] Sunter, A. B., "A statistical approach to record linkage; record linkage in medicine," *Proceedings of the International Symposium*, Oxford, July 1967; E. & S. Livingstone Ltd., London, 1968.
[10] Sunter, A. B., and Fellegi, I. P., "An optimal theory of record linkage," 36th Session of the International Statistical Institute, Sydney, Australia, 1967.
[11] Tepping, B. J., "Study of matching techniques for subscriptions fulfillment," National Analysts Inc., Philadelphia, August, 1955.
[12] Tepping, B. J., "A model for optimum linkage of records," *Journal of the American Statistical Association*, 63 (1968) 1321–1332.
[13] Tepping, B. J., and Chu, J. T., "A report on matching rules applied to readers digest data," National Analysts Inc., Philadelphia, August, 1958.

APPENDIX I

A FUNDAMENTAL THEOREM FOR RECORD LINKAGE

We stated that (μ, λ) is an admissible pair of error levels provided μ and λ are not both too large. We will make this statement more precise.

Let

$$U_n = \sum_{i=1}^{n} u_i; \qquad n = 1, 2, \cdots, N_\Gamma \tag{1}$$

$$U_0 = 0 \tag{2}$$

$$M_{n'} = \sum_{i=n'}^{N_\Gamma} m_i; \qquad n' = 1, 2, \cdots, N_\Gamma \tag{3}$$

$$M_{N_\Gamma+1} = 0 \tag{4}$$

and define $f(\mu)$, as shown in Figure 1, on the interval (0, 1) as the monotone decreasing polygon line passing through the points (U_n, M_{n+1}) for $n=0, 1, \cdots, N$. It is possible of course to state the definition more precisely, but unnecessary for our purposes.

The area contained by the axes and including the line $\lambda = f(\mu)$ defines the region of admissible pairs (μ, λ). In other words (μ, λ) is an admissible pair if

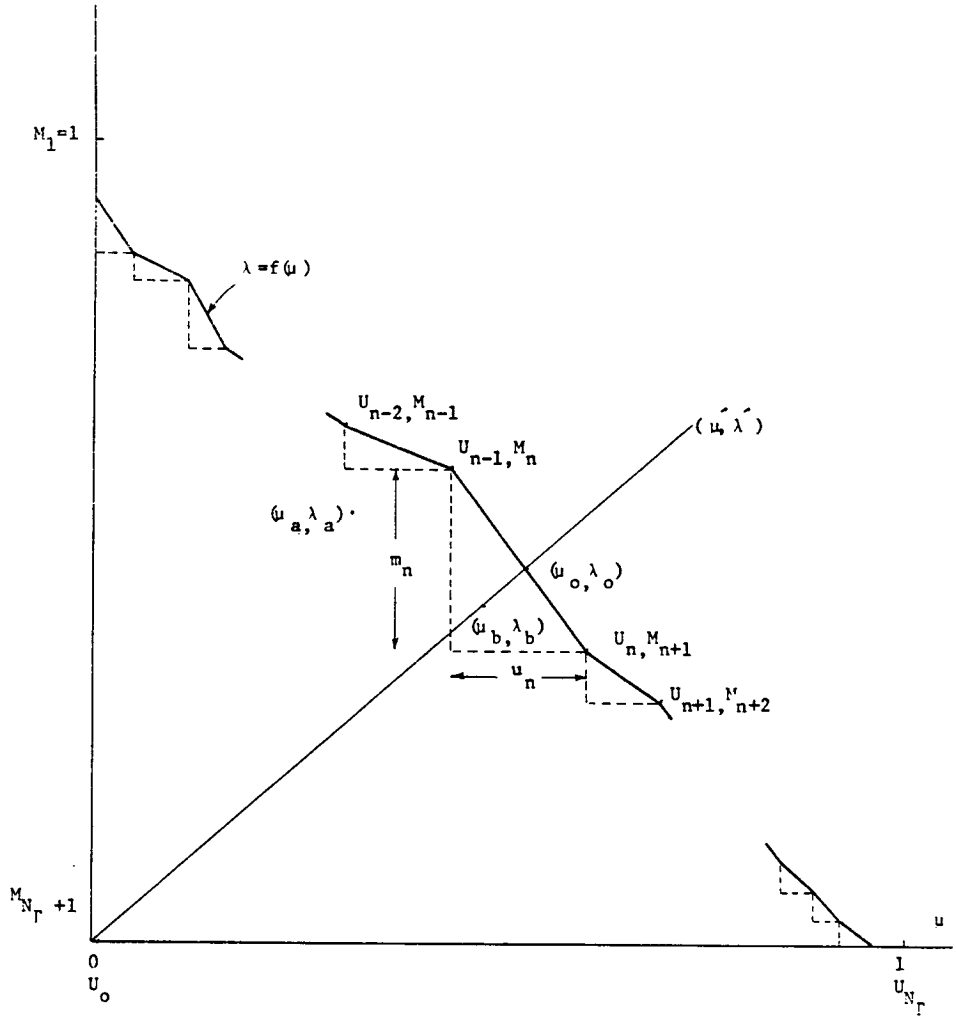

FIG. 1

$$0 < \lambda \leq f(\mu)$$
$$\text{and} \quad 0 < \mu. \tag{5}$$

Let $n(\mu)$ be the integer such that

$$U_{n(\mu)-1} < \mu \leq U_{n(\mu)} \tag{6}$$

and $n'(\lambda)$ the integer such that

$$M_{n'(\lambda)} \geq \lambda > M_{n'(\lambda)+1}. \tag{7}$$

Define

$$P_\lambda = \frac{\lambda - M_{n'(\lambda)+1}}{m_{n'(\lambda)}} \tag{8}$$

and

$$P_\mu = \frac{\mu - U_{n(\mu)-1}}{u_{n(\mu)}} . \tag{9}$$

It follows from the way in which the configurations were ordered and the restrictions on μ and λ that the denominators of the expressions on the right of (8) and (9) are positive.

It is easy to see from Figure 1 that

$$0 < P_\lambda \leq 1 \quad \text{and} \quad 0 < P_\mu \leq 1. \tag{10}$$

It is also clear from Figure 1 that (μ, λ) are admissible if and only if

$$\begin{aligned} &\text{(a)} \quad n'(\lambda) \geq n(\mu) + 1 \\ &\qquad \text{(e.g. } (\mu_a, \lambda_a) \text{ in Figure 1)} \\ &\qquad\qquad \text{or} \\ &\text{(b)} \quad n'(\lambda) = n(\mu) \quad \text{and} \quad P_\lambda + P_\mu \leq 1 \\ &\qquad \text{(e.g. } (\mu_b, \lambda_b) \text{ in Figure 1).} \end{aligned} \tag{11}$$

Thus (a) and (b) simply divide the admissible region into two areas, one bounded by the axes and the broken lines in Figure 1, and the other bounded by the broken lines and the polygon line $\lambda = f(\mu)$.

Finally, from Figure 1 and the definitions of $n(\mu)$ and $n'(\lambda)$ we see that $\lambda = f(\mu)$ if and only if

$$\begin{aligned} &\text{(a)} \quad n'(\lambda) = n(\mu) + 1 \quad \text{and} \quad P_\lambda = P_\mu \\ &\qquad \text{(i.e. the vertices of } \lambda = f(\mu)). \\ &\qquad\qquad \text{or} \end{aligned} \tag{12}$$

$$\begin{aligned} &\text{(b)} \quad n'(\lambda) = n(\mu) \quad \text{and} \quad P_\lambda + P_\mu = 1 \\ &\qquad \text{(i.e. points on } \lambda = f(\mu) \text{ other than vertices).} \end{aligned} \tag{13}$$

Let (μ, λ) be an admissible pair of error levels on Γ. We define a linkage rule $L_0(\mu, \lambda, \Gamma)$ as follows:

1) If $n'(\lambda) > n(\mu) + 1$ then

$$d_0(\gamma_i) = \begin{cases} (1, 0,) & \text{if } i \leq n(\mu) - 1 \\ (P_\mu, 1 - P_\mu, 0) & \text{if } i = n(\mu) \\ (0, 1, 0) & \text{if } n(\mu) + 1 \leq i \leq n'(\lambda) - 1 \\ (0, 1 - P_\lambda, P_\lambda) & \text{if } i = n'(\lambda) \\ (0, 0, 1) & \text{if } i \geq n'(\lambda) + 1 \end{cases}$$

2) If $n'(\lambda) = n(\mu)$ and $P_\lambda + P_\mu \leq 1$

$$d_0(\gamma_i) = \begin{cases} (1, 0, 0) & \text{if } i \leq n(\mu) - 1 \\ (P_\mu, 1 - P_\mu - P_\lambda, P_\lambda) & \text{if } i = n(\mu) = n'(\lambda) \\ (0, 0, 1) & \text{if } i \geq n'(\lambda) + 1. \end{cases}$$

(It is easy to see that (μ, λ) is admissible if and only if one of the two conditions above holds.)

We have now defined a linkage rule for an arbitrary pair of admissible levels (μ, λ). It follows immediately from the definition of $L_0(\mu, \lambda, \Gamma)$ that $P(A_2)=0$ if and only if $\lambda=f(\mu)$

Theorem: If (μ, λ) is an admissible pair of error levels on Γ then $L_0(\mu, \lambda, \Gamma)$ is the best linkage rule on Γ at the levels μ and λ. If (μ, λ) is not admissible on Γ then there are levels (μ_0, λ_0) with

$$\mu_0 \leqq \mu, \quad \text{and} \quad \lambda_0 \leqq \lambda \tag{14}$$

(with at least one of the inequalities in (14) being a definite inequality) such that $L_0^*(\mu_0, \lambda_0, \Gamma)$ is better than $L_0(\mu, \lambda, \Gamma)$ and for which

$$P_{L_0}(A_2) = 0. \tag{15}$$

This theorem explains the terminology "inadmissible." This simply means that we should not consider linkage rules at inadmissible error levels, since in this case L_0^* always provides a linkage rule at lower error levels for which we still have $P(A_2)=0$ (i.e. only the positive dispositions A_1 and A_3 occur).

Proof:

Let $L'(\mu, \lambda, \Gamma)$ be any linkage rule with admissible levels (μ, λ). Then $L'(\mu, \lambda, \Gamma)$ can be characterized by the set of decision functions

$$d'(\gamma_i) = (P'_{i1}, P'_{i2}, P'_{i3}), \qquad \sum_{j=1}^{3} P'_{ij} = 1 \qquad i = 1, 2, \cdots, N_\Gamma \tag{16}$$

where

$$P'_{ij} = P(A_j \mid \gamma_i), \qquad j = 1, 2, 3; \qquad i = 1, 2, \cdots, N_\Gamma. \tag{17}$$

Clearly

$$P_{L'}(A_1 \mid U) = \sum_{i=1}^{N_\Gamma} u_i P'_{i1} = \mu \tag{18}$$

$$P_{L'}(A_3 \mid M) = \sum_{i=1}^{N_\Gamma} m_i P'_{i3} = \lambda. \tag{19}$$

Consider the linkage rule $L_0(\mu, \lambda, \Gamma)$. It is characterized by equations analogous to (16) to (19) but P'_{ij} replaced by P_{ij} as defined above. We shall prove that

$$P(A_2 \mid L_0) \leqq P(A_2 \mid L') \tag{20}$$

According to the construction of L_0 the u_i which happen to be zero have the smallest subscripts, the m_i which happen to be zero have the largest subscripts. More rigorously, there are subscripts r and s such that

$$u_i = 0 \quad \text{if } i \leqq r-1, \qquad u_i > 0 \quad \text{if } i \geq r \tag{21}$$

$$m_i = 0 \quad \text{if } i \geqq s+1, \qquad m_i > 0 \quad \text{if } i \leq s \tag{22}$$

We have seen previously that

$$u_{n(\mu)} > 0$$

and

$$m_{n'(\lambda)} > 0$$

hence

$$n(\mu) \geqq r$$
$$n'(\lambda) \leqq s$$

hence

$$P_{i1} = 1 \qquad \text{for } i = 1, 2, \cdots, r-1 \tag{23}$$
$$P_{i3} = 1 \qquad \text{for } i = s+1, s+2, \cdots, N_\Gamma \tag{24}$$

that is, whenever u_i is zero then $P_{i1}=1$ and whenever $m_i=0$ then $P_{i3}=1$.

By definition of μ, it follows that

$$\sum_{i=1}^{N_\Gamma} u_i P_{i1} = \sum_{i=1}^{N_\Gamma} u_i P'_{i1} = \mu. \tag{25}$$

Putting $n=n(\mu)$ and observing that $P_{i1}=1$ if $i \leqq n-1$ we can express (25) as follows:

$$\sum_{i=1}^{n-1} u_i + u_n P_\mu = \sum_{i=1}^{N_\Gamma} u_i P'_{i1}$$

or

$$\sum_{i=1}^{n-1} u_i(1 - P'_{i1}) + u_n(P_\mu - P'_{n,1}) = \sum_{i=n+1}^{N_\Gamma} u_i P'_{i1}. \tag{26}$$

With the possible exception of the last term on the left it is clear that every term in (26) is non-negative. We assume, without loss of generality, that the term in question *is* non-negative for, if it were negative, we would simply transfer it to the other side of the equality and all of the steps to follow would hold. It follows that if not every term in (26) is equal to zero then both sides are positive. Assume for the moment that this is the case.

It follows from the ordering of Γ that

$$u_i m_j \leqq u_j m_i \qquad \text{whenever } i < j. \tag{27}$$

It is now seen that

$$\left[\sum_{j=n+1}^{N_\Gamma} m_j P'_{j1}\right]\left[\sum_{i=1}^{n-1} u_i(1 - P'_{i1}) + u_n(P_\mu - P'_{n,1})\right] \leqq \left[\sum_{i=1}^{n-1} m_i(1 - P'_{i1}) + m_n(P_\mu - P'_{n,1})\right]\left[\sum_{j=n+1}^{N_\Gamma} u_j P'_{j1}\right] \tag{28}$$

since by (27) every term in the expansion of the left hand side is of the form

$$m_j u_i P'_{j1}(1 - P'_{i1}) \quad \text{or} \quad m_j u_n P'_j(P_\mu - P_{n,1}) \qquad (i \leqq n < j)$$

and corresponding to each there is a similar term on the right hand side but with $m_j u_i$ replaced by $m_i u_j$ and $m_j u_n$ replaced by $m_n u_j$. Dividing (28) by (26) we get

$$\sum_{j=n+1}^{N\Gamma} m_j P'_{j1} \leqq \sum_{j=1}^{n-1} m_j(1 - P'_{j1}) + m_n(P_\mu - P'_{n,1})$$

or

$$\sum_{i=1}^{N\Gamma} m_i P'_{i1} \leqq \sum_{i=1}^{N\Gamma} m_i P_{i1}. \tag{29}$$

If every term in (26) was zero (29) would still hold since in that case we would have

$$P_{i1} = P'_{i1} \qquad \text{for } i \geqq r$$

i.e. whenever $u_i \neq 0$ and we would have

$$P_{i1} = 1 \geqq P'_{i1} \qquad \text{for } i \leqq r - 1$$

because of (23) and because $P'_{i1} \leqq 1$ for every i. Hence (29) would hold in this case as well.

By definition

$$\sum_{i=1}^{N\Gamma} m_i P'_{i3} = \sum_{i=1}^{N\Gamma} m_i P_{i3} = \lambda. \tag{30}$$

From (29) and (30) we get

$$\sum_{i=1}^{N\Gamma} m_i(P'_{i1} + P'_{i3}) \leqq \sum_{i=1}^{N\Gamma} (P_{i1} + P_{i3})$$

or

$$\sum_{i=1}^{N\Gamma} m_i(1 - P'_{i2}) \leqq \sum_{i=1}^{N\Gamma} m_i(1 - P_{i2}). \tag{31}$$

Because

$$\sum_{i=1}^{N\Gamma} m_i = 1, \qquad \text{we get}$$

$$\sum_{i=1}^{N\Gamma} m_i P_{i2} \leqq \sum_{i=1}^{N\Gamma} m_i P'_{i2}$$

or

$$P_{L_0}(A_2 \mid M) \leqq P_{L'}(A_2 \mid M). \tag{32}$$

It can be shown similarly that

$$P_{L_0}(A_2 \mid U) \leqq P_{L'}(A_2 \mid U). \tag{33}$$

But (32) and (33) together state that

$$P(A_2 \mid L_0) \leqq P(A_2 \mid L') \tag{34}$$

which completes the proof of the first part of the theorem. Note that we have actually proved more than (34) since we have proved that L_0 is optimal separately under both the conditions M *and* the condition U. This also explains why the prior probabilities $P(M)$ and $P(U)$ do not enter either the statement or the proof of the theorem; our result is independent of these prior probabilities. The underlying reason, of course, lies in the fact that the error levels are concerned with conditional probabilities of misallocation. The situation would change if one tried to minimize the unconditional probability of misallocation or if one tried to minimize some general loss function.

As for the proof of the second part, let (μ', λ') be an inadmissible pair of error levels $(0<\mu<1,\ 0<\lambda<1)$. Since $f(\mu)$ is a strictly monotone decreasing continuous function in the range determined by

$$0 < \mu < 1$$
$$0 < f(\mu) < 1$$

it will intersect at a unique point the straight line drawn through (0, 0) and (μ', λ'). This is illustrated in Figure 1. Denote this point by (μ_0, λ_0). Then

$$0 < \mu_0 < \mu' < 1$$
$$0 < \lambda_0 < \lambda' < 1$$

and

$$\lambda_0 = f(\mu_0). \tag{35}$$

The linkage rule $L_0(\mu_0, \lambda_0, \Gamma)$ is, in light of (36), (12), and (13) such that

$$P(A_2 \mid L_0) = 0.$$

Hence $L_0(\mu_0, \lambda_0, \Gamma)$ is a better linkage rule than any other linkage rule at the level (μ', λ').

This completes the full proof of our theorem.

The form of the theorem given in the text is an immediate corollary of the theorem above and the expression (11).

APPENDIX II

METHOD II FOR THE CALCULATION OF WEIGHTS

Denoting

$$N_A N_B = c$$

the equations resulting from (37) to (39) by dropping expected values can be written as

$$M_k = \frac{N}{c} \prod_{j=1, j\neq k}^{3} m_j + \frac{c - N}{c} \prod_{j=1, j\neq k}^{3} u_j \qquad k = 1, 2, 3 \tag{1}$$

$$U_k = \frac{N}{c} m_k + \frac{c - N}{c} u_k \qquad k = 1, 2, 3 \tag{2}$$

$$M = \frac{N}{c} \prod_{j=1}^{3} m_j + \frac{c - N}{c} \prod_{j=1}^{3} u_j. \tag{3}$$

We introduce the transformation

$$m_k^* = m_k - U_k \tag{4}$$

$$u_k^* = u_k - U_k. \tag{5}$$

Substituting m_k and u_k from (4) and (5) into (2) we obtain

$$\frac{N}{c} m_k^* + \frac{c - N}{c} u_k^* = 0 \qquad k = 1, 2, 3. \tag{6}$$

Substituting (4) and (5) into (1) and then substituting in the resulting equations u_k^* from (6) we obtain

$$\prod_{j=1, j\neq k}^{3} m_j^* = \frac{c - N}{N} \left[M_k - \prod_{j=1, j\neq k}^{3} U_j \right] \qquad k = 1, 2, 3. \tag{7}$$

Denoting

$$R_k = M_k - \prod_{j=1, j\neq k}^{3} U_j \qquad k = 1, 2, 3 \tag{8}$$

we obtain by multiplying the three equations under (7) and by taking square roots

$$\prod_{j=1}^{3} m_j^* = \left(\frac{c - N}{N} \right)^{\frac{1}{2}} \left(\prod_{j=1}^{3} R_j \right)^{\frac{1}{2}} \tag{9}$$

Dividing (9) by (7) and putting

$$X = \sqrt{(c - N)/N} \tag{10}$$

$$B_k = \sqrt{\prod_{j=1, j\neq k}^{3} R_j / R_k} \qquad k = 1, 2, 3 \tag{11}$$

we get

$$m_k^* = B_k X \qquad k = 1, 2, 3 \tag{12}$$

and, from (4) to (6),

$$m_k = U_k + B_k X \qquad k = 1, 2, 3 \tag{13}$$

$$u_k = U_k - B_k / X \qquad k = 1, 2, 3. \tag{14}$$

We can now substitute into (3) m_k and u_k from (13) and (14) respectively and N as expressed from (10). We obtain

$$\frac{1}{X^2 + 1} \prod_{j=1}^{3} (U_j + B_j X) + \frac{X^2}{X^2 + 1} (U_j - B_j / X) = M. \tag{15}$$

After expanding (15), some cancellations and substitution of B_k from (11) we get the following quadratic equation in X:

$$\sqrt{\prod_{j=1}^{3} R_j}\,(X^2 - 1) + \left[\prod_{j=1}^{3} U_j + \sum_{j=1}^{3} R_j U_j - M\right] X = 0. \tag{16}$$

The positive root of this equation is

$$X = \left\{M - \sum_{j=1}^{3} R_j U_j - \prod_{j=1}^{3} U_j + \sqrt{\left[M - \sum_{j=1}^{3} R_j U_j - \prod_{j=1}^{3} U_j\right]^2 + 4 \prod_{j=1}^{3} R_j}\right\} / 2 \sqrt{\prod_{j=1}^{3} R_j}. \tag{17}$$

The estimates of m_k, u_k and N are now easily obtained from (10), (13) and (14).

Having solved these equations we can proceed to estimate the specific values of $m(\gamma)$ and $u(\gamma)$ which are required. We introduce some additional notation which, as before, refers to observable frequencies:

$M_k(\gamma_i^k)$ = the proportion of "agreement" in all components except the kth; the specific configuration γ_i^k in the kth component

$U_1(\gamma_i^2)$ = the proportion of "agreement" in the first, γ_i^2 in the second and any configuration in the third component

$U_1(\gamma_i^3)$ = the proportion of "agreement" in the first, γ_i^3 in the third and any configuration in the third component

$U_2(\gamma_i^1)$ = the proportion of γ_i^1 in the first, "agreement" in the second and any configuration in the third component.

The required values of $m(\gamma_i^k)$ and $u(\gamma_i^k)$ are estimated as

$$m(\gamma_i^1) = \frac{M_1(\gamma_i^1) - u_3 U_2(\gamma_i^1)}{m_2(m_3 - u_3)}(X^2 + 1) \tag{18}$$

$$m(\gamma_i^2) = \frac{M_2(\gamma_i^2) - u_3 U_1(\gamma_i^2)}{m_1(m_3 - u_3)}(X^2 + 1) \tag{19}$$

$$m(\gamma_i^3) = \frac{M_3(\gamma_i^3) - u_2 U_1(\gamma_i^3)}{m_1(m_2 - u_2)}(X^2 + 1) \tag{20}$$

$$u(\gamma_i^1) = \frac{m_3 U_2(\gamma_i^1) - M_1(\gamma_i^1)}{u_2(m_3 - u_3)} \frac{X^2 + 1}{X^2} \tag{21}$$

$$u(\gamma_i^2) = \frac{m_3 U_1(\gamma_i^2) - M_2(\gamma_i^2)}{u_1(m_3 - u_3)} \frac{X^2 + 1}{X^2} \tag{22}$$

$$u(\gamma_i^3) = \frac{m_2 U_1(\gamma_i^3) - M_2(\gamma_i^3)}{u_1(m_2 - u_2)} \frac{X^2 + 1}{X^2} \tag{23}$$

The formulae (18) to (23) are easily verified by expressing the expected values of the quantitites $M_k(\gamma_i^k)$, $U_1(\gamma_i^2)$, etc. in terms of m_k, u_k, $m(\gamma_i^k)$ and $u(\gamma_i^k)$,

dropping the expected values and solving the resulting equations (there will be two equations for each pair $m(\gamma_i^k)$ and $u(\gamma_i^k)$).

The necessary and sufficient conditions for the mechanical validity of the formulae in this section are that

$$m_k \neq u_k \qquad k = 1, 2, 3$$

and

$$R_k > 0 \qquad k = 1, 2, 3$$

Since

$$m_k = m(S_k) = \Pr(S_k \mid M)$$
$$u_k = u(S_k) = \Pr(S_k \mid U)$$

clearly for sensible definitions of "agreement" $m_k > u_k$ should hold for $k = 1, 2, 3$. In this case $R_k > 0$ will hold as well. The latter statement can easily be verified by substituting (1) and (2) into (8).