Assessing disclosure risk for record linkage

Book section

Assessing Disclosure Risk for Record Linkage

Chris Skinner

Southampton Statistical Sciences Research Institute
University of Southampton
Southampton SO17 1EF, United Kingdom

Abstract. An intruder seeks to match a microdata file to an external file using a record linkage technique. The identification risk is defined as the probability that a match is correct. The nature of this probability and its estimation is explored. Some connections are made to the literature on disclosure risk based on the notion of population uniqueness.

Keywords: identification; log-linear model; match; misclassification; uniqueness

1 Introduction

Statistical agencies are obliged to protect confidentiality when they release outputs. One potential threat to confidentiality is the use of record linkage methods [1, 2, 3]. The concern is that an 'intruder' might link an element of an agency's output to a known individual (or other unit) in some external data source and, if the link is correct, succeed in identifying an individual who provided data upon which the output is based. Such identification (identity disclosure) might lead to the disclosure of further information about this individual.

This threat is most natural to consider when the output consists of a microdata file. In this paper we suppose the agency releases a file containing records for a sample of units, with each record containing the values of various variables. These values may have been masked by statistical disclosure control (SDC) methods, although we suppose there remains a one to one correspondence between the records and the units which provided the data. Thus, identification of these units could, in principle, occur via record linkage to an external file of known units. We suppose that linkage takes place by matching the values of a subset of the variables, 'key variables', shared between the microdata and the external file.

The main aim of this paper is to consider approaches to measuring and estimating the risk of identity disclosure in this setting. A secondary aim is to link this work with other approaches in the literature to assessing identification risk which have centred on concerns about the existence of 'population uniques', i.e. records which are unique in the population with respect to their values of the key variables.

Possibly the earliest contribution to assessing the identification risk arising from record linkage is by Spruill in [4]. She considers linkage methods which match by minimizing a distance measure and combines the definition of risk with the method for assessing it. The approach is based upon a re-identification experiment where each

record in a microdata file, which has been masked by an SDC method, is matched to the original unmasked file and the closest record in the latter file selected. The risk is defined essentially as the proportion of such matches which are correct. She also notes that account might be taken of 'near matches'. This broad approach has been adopted or discussed in much subsequent literature, e.g. [5, 1, 6, 7].

There are, however, some problems with using the empirical proportion of correct matches as a measure of risk. First, the original unmasked file is acting as a surrogate for an external file held by the intruder in such approaches. The use of this file represents a highly conservative approach to risk assessment since it ignores the protective effect of sampling and, even if there are some common units in the microdata and external files, the values of the variables for these units in the two files are likely to differ for many practical reasons e.g. differences in measurement. To address this concern, the original unmasked file might be replaced by an alternative surrogate external file constructed by the agency. For example, it is reported in [8] that the US National Center for Education Statistics uses certain commercially available school files. Agencies may also consider using other datasets which they collect (from other surveys) or constructing synthetic files from the original unmasked file which take account of sampling and measurement error.

A second more conceptual problem with this approach is that it can fail to reflect adequately the information available to the intruder. Suppose, for example, that the overall proportion of correct matches is 5% and that the agency considers this sufficiently low. Suppose, however, that the intruder could determine which 5% of his claimed matches are correct and which 95% are incorrect. Then the intruder could claim some matches with 100% confidence and this might be deemed an unacceptable disclosure risk. On the other hand, suppose the agency chooses to calculate its proportions separately according to different areas and observes that the proportions vary across areas from 0% to 70%. It might deem the release of data for those areas with proportions as high as 70% as unacceptable. However, if the intruder could only determine that the overall rate of a correct match was 5% (in practice, the intruder will have difficulty determining the proportion of correct matches since it requires knowledge of the true identities of the records in the microdata, information unavailable to the intruder) and was unable to identify areas where it was higher, the agency's judgment would be over-conservative.

In this paper we suppose that it is necessary for the intruder to have evidence that the link is 'likely' to be correct. Identification risk is defined as the probability that a match is correct, conditional on data assumed available to the intruder, c.f. [9, 10], and it is required that this probability can be estimated reliably from these data. We suppose that the agency might use empirical proportions of correct matches as a means of validating these estimates but not as a direct means of estimation.

We focus in this paper on probabilistic record linkage methods (based on the approach of Fellegi and Sunter in [11] (hereafter referred to as FS) rather than methods based on distance measures. These probabilistic methods are most naturally adapted to assess the probability of a correct match. Indeed, part of conventional record linkage methodology is the estimation of false match rates and one might, as a first approach, take one minus the estimated false match rate as a measure of identification risk. However, in conventional applications of record linkage, incorrect matches (false positives or false negatives) are only of interest because of their

statistical consequences for samples as a whole. FS (p. 1196) state that 'we are not concerned with the *probability* of [these two kinds of erroneous matches]…but rather with the *proportion* of occurrences of these two events in the long run'. In contrast, requirements to protect the confidentiality of every individual imply that an agency may be interested in the probability of a correct match for a single individual.

The paper is organized as follows. First, a framework for the use of record linkage for identification is set out in Section 2. Expressions for the probability of a correct match are obtained in Section 3. After briefly considering issues relating to key variables in Section 4, the estimation of the probability of a correct match is considered in Section 5.

2 The Use of Record Linkage to Achieve Identification

Consider a survey microdata file containing records for a sample of responding units s_1 drawn from a finite population P. Each record will include variables needed by genuine users of the file, but is supposed not to include directly identifying variables like name and address. Suppose an intruder has access to this file and wishes to identify one or more units in s_1. The intruder matches the file to an external file of records for another sample of units $s_2 \subset P$, for which the identities are known and for which it is feasible that the intersection $s_{12} = s_1 \cap s_2$ is non-empty. (We assume that the definition of the population P is public and that the intruder can thus remove any records in the external file which do not belong to P – hence we do not need to allow for s_1 and s_2 to be drawn from different populations.)

Suppose matching is based upon the values of variables, which appear in both files: the *key variables* [12]. Let $\tilde{X}_a$ denote the value of the vector of key variables for unit a in the microdata ($a \in s_1$) and X_b the corresponding value for unit b in the external database ($b \in s_2$). The difference in notation between $\tilde{X}$ and X allows for the possibility that the variables are recorded in a different way in the two data sources. This difference might arise from various reasons, including measurement error (in either source) or the application of a perturbative SDC method to the microdata file. Following FS, suppose the intruder undertakes linkage by calculating a comparison vector $\gamma(\tilde{X}_a, X_b)$ for pairs of units $(a,b) \in s_1 \times s_2$, where the function $\gamma(.,.)$ takes values in some finite comparison space Γ.

Example 1: Exact Matching on Categorical Key Variables

Suppose $\tilde{X}$ and X take only K possible values, denoted $\{1,...,K\}$ without loss of generality. Let $\Gamma = \{1,2,...,K+1\}$ and define the comparison vector by $\gamma(\tilde{X}, X) = j$ if $\tilde{X} = X = j$, $j = 1,2,...,K$, $\gamma(\tilde{X}, X) = K+1$ otherwise. In this case, an intruder might consider any pair $(a,b) \in s_1 \times s_2$ for which $\gamma(\tilde{X}_a, X_b) \leq K$ as a potential match, but rule out of consideration any pair for which $\gamma(\tilde{X}_a, X_b) = K+1$.

Suppose the intruder seeks to use the comparison vectors to identify one or more pairs $(a,b) \in s_1 \times s_2$ which contain identical units, i.e. are of the form (a,a) where $a \in s_{12}$. Since the number of pairs in $s_1 \times s_2$ may be very large, the intruder may only consider pairs which fall in a set $\tilde{s} \subset s_1 \times s_2$. Partition $\tilde{s}$ into $M = \{(a,b) \in \tilde{s} \mid a = b, a \in s_{12}\}$, the pairs of common units, and $U = \{(a,b) \in \tilde{s} \mid a \in s_1, b \in s_2, a \neq b\}$, the pairs of different units. The problem faced by the intruder is how to use comparison vector values to classify pairs from $\tilde{s}$ into M or U. An optimum strategy is shown by FS to be based upon a comparison of the probability distributions of the comparison vector between M and U, i.e. a comparison of

$$m(\gamma) = \Pr[\gamma(\tilde{X}_a, X_b) = \gamma \mid (a,b) \in M], \tag{1}$$

and $$u(\gamma) = \Pr[\gamma(\tilde{X}_a, X_b) = \gamma \mid (a,b) \in U], \quad \gamma \in \Gamma. \tag{2}$$

We discuss the nature of these probabilities in the next section. FS show that an optimal approach for the intruder is to order pairs in $\tilde{s}$ according to the likelihood ratios $m(\gamma)/u(\gamma)$, treating pairs with higher values of this ratio as more likely to belong to M. Our aim is to explore the probability of a correct match for pairs selected in this way.

3 The Probability of a Correct Match

Given a pair (a,b), linked using its value of the comparison vector as described after (1) and (2), the probability that the pair represents a correct match, that is $a = b$, may be defined as $p_{M|\gamma} = \Pr[(a,b) \in M \mid \gamma(\tilde{X}_a, X_b)]$, i.e. the conditional probability that the pair is in M given that it is in $\tilde{s}$ and that the comparison vector takes the value γ. To express $p_{M|\gamma}$ in terms of $m(\gamma)$ and $u(\gamma)$, let:

$$p = \Pr[(a,b) \in M], \tag{3}$$

be the probability that the pair is in M given that it is in $\tilde{s}$ and, using Bayes theorem, we obtain

$$p_{M|\gamma} = m(\gamma)p / [m(\gamma)p + u(\gamma)(1-p)] \quad . \tag{4}$$

Sorting pairs according to this 'posterior' probability is equivalent to sorting according to the likelihood ratio $m(\gamma)/u(\gamma)$. From the SDC perspective, expression

(4) may be interpreted as the identification risk for a pair (a,b), i.e. the probability that a and b are identical, given the value of the comparison vector. From the record linkage perspective, expression (4) is the probability of a correct match or alternatively one minus the probability of a false match [13].

Expressions (1), (2) and (3) are, of course, dependent on the way the probabilities are defined. Our basic approach in this paper is to suppose that the probabilities are defined with respect to the following three processes:

(i) a random selection (with equal probability) of the pair (a,b) from $\tilde{s} = M \cup U$;

(ii) a random process of generating $\tilde{X}_a$;

(iii) a specified probability design for the selection of s_1 from P ;

where the population P and the values X_a for units in the population are treated as fixed. Evaluating the probabilities over (i), holding s_1 and the $\tilde{X}_a$ fixed, we may write

$$m(\gamma) = E[n_{M\gamma} / n_M] \ , \ u(\gamma) = E[n_{U\gamma} / n_U] \ , \tag{5}$$

where n_M and n_U are the numbers of pairs in M and U respectively, $n_{M\gamma}$ and $n_{U\gamma}$ are the corresponding numbers of these pairs for which the comparison vector takes the value γ and the expectation is with respect to (ii) and (iii). We may thus interpret $m(\gamma)$ and $u(\gamma)$ as the expected relative frequencies of the different comparison vectors within M and U respectively. Similarly, we may write

$$p = E(n_M / \tilde{n}) \, , \tag{6}$$

where $\tilde{n}$ is the number of pairs in $\tilde{s}$ and the expectation is with respect to (iii). To explore the form of $p_{M|\gamma}$ further under (i), (ii) and (iii), consider two special cases.

Example 1(continued) Exact matching with no misclassification

Suppose exact matching is used as defined earlier and that: $\tilde{X}_a = X_a$ for all units $a \in P$ (i.e. no misclassification); $s_2 = P$ and $\tilde{s} = s_1 \times s_2$. Let $n_1 = |s_1|$ and $N = |P|$. Noting that $n_M = n_1$ and $\tilde{n} = n_1 N$, we obtain from (5) and (6):

$$m(j) = E[f_j / n_1] \ , \qquad u(j) = E\left(\frac{f_j (F_j - 1)}{n_1 (N-1)} \right) \ , \qquad j = 1, \ldots, K$$

$$p = E[n_1 / (n_1 N)] = 1/N \, , \tag{7}$$

where f_j and F_j are the numbers of units with $X_a = j$ in s_1 and P respectively. Using Bayes theorem we obtain:

$$\Pr[(a,b) \in M \mid \gamma(\tilde{X}_a, X_b) = j] = 1/F_j \quad . \tag{8}$$

This result if free of any assumptions about the sampling scheme. Expression (8) is familiar in the disclosure risk literature, e.g. [14]. It is common to argue, however, that agencies should design release strategies so that an intruder could not know the value of F_j from external information [10]. Note that, in particular, this requires assuming that $s_2 \neq P$. Otherwise, the intruder could determine F_j from knowledge of X_a for $a \in P$. If F_j is unknown to the intruder, the uncertainty about F_j needs to be integrated out of the expression for the identification risk, subject to conditioning on the information available to the intruder. This integration is most naturally done by revising the probability mechanisms (i)-(iii) above to include a process which generates the values X_a for units in the population. Under this extended probability mechanism, the identification risk becomes $E(1/F_j \mid data)$, where *data* represents the data available to the intruder. We shall return to this issue in Section 5. First, we extend the result in (8) to the case when $\tilde{X}_a$ may be derived from X_a by a process of misclassification and s_2 may be any proper subset of P.

Example 1 (continued) Exact matching with misclassification

Suppose again that exact matching is used and that $\tilde{s} = s_1 \times s_2$. We now allow s_2 to be any proper subset of P and suppose that each $\tilde{X}_a$ is determined from X_a as follows

$$\Pr(\tilde{X}_a = j \mid X_a = k) = \theta_{jk} \text{ , for all } a \in P \text{ ,} \qquad (9)$$

where θ_{jk} is an element of a misclassification matrix with columns which sum to 1. We now obtain

$$m(j) = E[f_j^{12} / n_{12}] \text{ , } u(j) = E\left(\frac{\tilde{f}_j f_j - f_j^{12}}{n_1 n_2 - n_{12}} \right), \qquad j = 1,...,K$$

$$p = E[n_{12} / (n_1 n_2)],$$

where f_j^{12} is the number of units in s_{12} with $X_a = j$ and $\tilde{X}_a = j$, $\tilde{f}_j$ is the number of units in s_1 with $\tilde{X}_a = j$ and f_j is the number of units in s_2 with $X_a = j$. If we suppose that Bernoulli sampling is employed with inclusion probability π we have $n_{12} \doteq n_2 n_1 / N$ so that $p \doteq 1/N$ and $n_1 n_2 - n_{12} \doteq (N-1) n_{12}$. It follows that

$$\Pr[(a,b) \in M \mid \gamma(\tilde{X}_a, X_b) = j] \doteq E\left(\frac{f_j^{12}}{\tilde{f}_j f_j} \right),$$

where the expectation is with respect to both the sampling and the misclassification mechanisms. We have $E(f_j^{12}) = \pi \theta_{jj} f_j$ and $E(\tilde{f}_j) = \pi \tilde{F}_j$, where $\tilde{F}_j$ is the number of units in P with $\tilde{X}_a = j$ (imagining that the misclassification takes place before the sampling). Hence we may write

$$\Pr[(a,b) \in M \mid \gamma(\tilde{X}_a, X_b) = j] \doteq \frac{\theta_{jj}}{\tilde{F}_j} \ . \tag{10}$$

Note that this expression applies for any choice of s_2, which may be selected arbitrarily. The expression in (4) for the probability of a correct match and the special cases in (8) and (10) apply to a pair of records (a,b) with a specific agreement pattern γ. This notion may be extended to apply to a class of pairs, $\hat{M}$, for which the likelihood ratio is above some threshold, say $\hat{M} = \{(a,b) \mid \gamma(\tilde{X}_a, X_b) \in \Gamma_M\}$, where Γ_M is the set of agreement patterns γ for which $m(\gamma)/u(\gamma)$ is above a threshold specified by the intruder as determining which pairs to declare as links.

A key issue for identification risk assessment is how to estimate $p_{M|\gamma}$ and, more specifically, how to estimate $p, m(\gamma)$ and $u(\gamma)$. We discuss this in section 5. Before then, we consider the record linkage approach further.

4 Taking Account of Key Variable Structure

In practice it is usual to base the comparison vector $\gamma(\tilde{X}_a, X_b)$ upon the separate comparisons of C key variables. Letting $\tilde{X} = (\tilde{X}^1, ..., \tilde{X}^C)$ and $X = (X^1, ..., X^C)$ we write

$$\gamma(\tilde{X}_a, X_b) = [\gamma^1(\tilde{X}_a^1, X_b^1), ..., \gamma^C(\tilde{X}_a^C, X_b^C)], \tag{11}$$

where $\gamma^c(\tilde{X}^c, X^c)$ denotes the comparison vector for the c^{th} key variable.

Example 2. Comparison vectors for simple agreements between continuous or categorical key variables, c.f. [15]

Let $\gamma^c(\tilde{X}^c, X^c) = 1$ if $\tilde{X}^c \sim X^c$ and $\gamma^c(\tilde{X}^c, X^c) = 0$, otherwise, $c = 1, 2, ..., C$, where $\sim$ is a specified agreement relation. Then

$\Gamma = \{(\gamma^1, \gamma^2, ..., \gamma^C) \mid \gamma^c = 0,1;\ c = 1, 2, ..., C\} = \{0,1\}^C$ and $|\Gamma| = 2^C$.

Example 3. Comparison vectors for agreements between categorical key variables

Suppose $\tilde{X}^c$ and X^c are categorical, taking values $j^c = 1, 2, ..., t^c$, and $\gamma^c(\tilde{X}^c, X^c) = j^c$ if $\tilde{X}^c = X^c = j^c$, $j^c = 1, 2, ..., t^c$, $\gamma^c(\tilde{X}^c, X^c) = t^c + 1$ otherwise, $c = 1, 2, ..., C$. Then

$\Gamma = \{(\gamma^1, \gamma^2, ..., \gamma^C) \mid \gamma^c = 1, ..., t^c + 1, c = 1, 2, ..., C\}$ and $|\Gamma| = \prod_{c=1}^{C} (t^c + 1)$.

Given the large potential size of Γ when C is at all large, it is common to restrict attention to a subspace Γ^* of Γ. A common approach is to partition the set of possible values of a specified subset of the key variables into blocks (e.g. [16]) so that the intruder only examines pairs for matching for which the values of these key variables fall in the same block. This constraint is typically equivalent to restricting attention to a subspace Γ^* of Γ.

The estimation of $m(\gamma)$ and $u(\gamma)$ is challenging if $|\Gamma|$ is large, as is likely in Examples 2 and 3 if C is at all large. It is therefore common to make simplifying assumptions, in particular, following FS, to treat the C agreement patterns in (11) as independent within M and U, i.e.

$$m(\gamma) = m_1(\gamma^1)m_2(\gamma^2)...m_C(\gamma^C) \text{ and } u(\gamma) = u_1(\gamma^1)u_2(\gamma^2)...u_C(\gamma^C) \quad (12)$$

where $m_c(\gamma^c) = \Pr[\gamma^c(\tilde{X}_a^c, X_b^c) = \gamma^c \mid (a,b) \in M]$ and $u_c(\gamma^c) = \Pr[\gamma^c(\tilde{X}_a^c, X_b^c) = \gamma^c \mid (a,b) \in U]$, $c = 1,2,...,C$. We refer to this assumption as *independence of agreement patterns***.** In the categorical variable case of Example 3 with misclassification defined as in (9), a sufficient condition for the independence of agreement patterns is that misclassification operates independently, variable by variable, and that the key variables are themselves independent.

5. Estimation

In this section we consider the estimation of the probability of a correct match, $p_{M|\gamma}$, defined in section 3. We assume that the estimator is a function only of data which is available to the intruder and thus rule out the possibility of using a training sample, c.f. [13]. In this case, one approach would be to use a *mixture model*, where $p, m(\gamma)$ and $u(\gamma)$ are treated as unknown parameters in a model for the observed values of the comparison vectors. The model is a mixture of models for M and U, treated as latent classes, and maximum likelihood estimation is used for parameter estimation (e.g. FS Method 2; [15, 17]). This modelling approach has found some success in record linkage applications where very strong identifying information, such as name and address, is available. On the other hand, it has been less successful when the distributions of the comparison vectors for M and U are not well-separated or are not each unimodal [15, 18] and this may be the case in practice in many SDC contexts, e.g. for social survey data. This is a matter for further empirical investigation, however, which we do not attempt in this paper.

Instead, we approach the estimation problem more directly by considering expressions for $p_{M|\gamma}$ in terms of our assumed probability mechanisms, as in section 3, and then considering how to estimate these expressions, from the data available to the intruder as well as possible additional external sources. This approach is analogous to Method 1 of FS. Since $p_{M|\gamma}$ is a function of $p, m(\gamma)$ and $u(\gamma)$, we also discuss the

problem of estimating these parameters to gain a better understanding of the general estimation problem. We first return to the two examples in Section 3.

Example : Exact matching with no misclassification

We obtained $p_{M|\gamma} = 1/F_j$ in expression (8) but argued, following this expression, that a more suitable measure will usually be $E(1/F_j \mid data)$. The evaluation of this conditional expectation is discussed in [19] under the assumption that the F_j are generated from a Poisson log-linear model and that the sample frequencies f_j represent the *data*. Treating the pairs (f_j, F_j) as independent, the conditional probability may then be expressed as $E(1/F_j \mid f_j)$ and a closed form expression may be obtained under the Poisson log-linear model and a Bernoulli sampling assumption. The conditional probability will be highest for cases which are unique in the sample, i.e. $f_j = 1$. The conditional probability may be estimated by estimating the log-linear model parameters and plugging these estimates into the expression for the conditional probability.

Example 1: Exact matching with misclassification

We obtained the approximate expression $p_{M|\gamma} \doteq \theta_{jj}/\tilde{F}_j$ in expression (10) . As above, we may argue that in practice $\tilde{F}_j$ will be unknown and a more suitable measure is $\theta_{jj}E(1/\tilde{F}_j \mid \tilde{f}_j)$. The second component of this expression, $E(1/\tilde{F}_j \mid \tilde{f}_j)$, may be estimated by applying the methodology of [19] to the observed microdata. The misclassification probability θ_{jj} might be estimated by making some approximating assumptions and using external evidence on the misclassification process. One assumption may be that some of the key variables are subject to no misclassification, as is commonly assumed for blocking variables, and that misclassification on the remaining variables is not dependent upon the values of such correctly classified variables. A further assumption may be that the remaining key variables are misclassified independently. This may be related to but is not the same as the earlier assumption of independence of agreement patterns. Under the independence of misclassification assumption, θ_{jj} may be expressed as a product of correct classification probabilities for the different key variables. This may need to be modified to allow for the possibility that the values of some key variables are missing.

To better understand the nature of the general estimation problem, now consider the separate estimation of $p, m(\gamma)$ and $u(\gamma)$. Consider p first. If $\tilde{n}$ is large we have from (6) that $p \doteq n_M/\tilde{n}$. The intruder knows the value of $\tilde{n}$ and so needs to estimate n_M in order to estimate p. We know $n_M \le n_{12}$, where $n_{12} = |s_{12}|$. And if we take the worst case, where the intruder selects $\tilde{s}$ in such a way that it includes all possible common pairs (i.e. all (a,a) where $a \in s_{12}$) then we have $n_M = n_{12}$. Thus, in order to estimate p, it suffices to estimate n_{12}. We *s*uppose the intruder can determine

inclusion probabilities $\pi_i = \Pr(i \in s_1)$ for $i \in s_2$. This is plausible. Often inclusion probabilities are equal in social surveys or else they will vary by strata which may be known for units in s_2. Since we have $n_{12} = E(\sum_{i \in s_2} \pi_i)$, where the expectation is with respect to the sampling scheme for s_1, the intruder can estimate n_{12} by $\hat{n}_{12} = \sum_{i \in s_2} \pi_i$ and hence estimate p by $\hat{p} = \hat{n}_{12} / \tilde{n}$. Note also that some adjustment will usually be necessary for nonresponse (e.g. by multiplying π_i by a response rate). Often in social surveys the inclusion probabilities π_i will be small and so $\hat{n}_{12}$ is only likely to be to have reasonable relative precision as an estimator if the size of the external database is large, representing a substantial proportion of the population. The extent to which p may be estimated reliably also, of course, depends upon this condition.

Let us now turn to the estimation of $m(\gamma)$ and $u(\gamma)$. Consider Example 1 with misclassification again, where we wish to estimate $m(\gamma)$ and $u(\gamma)$ for $j = 1,...,K$. We may write $m(j) = \theta_{jj} E[n_{12j} / n_{12}]$, where n_{12j} is the number of units in s_{12} with $\gamma = j$. And under Bernoulli (or equal probability) sampling we may write $E[n_{12j} / n_{12}] = f_j / n_2$, so that $m(j) = \theta_{jj} f_j / n_2$. And to first approximation (Jaro, 1989) we have: $u(j) \doteq (\tilde{f}_j / n_1)(f_j / n_2)$. The right hand side of this expression provides an estimator of $u(j)$ which should be reliable when $\tilde{f}_j$ and f_j are not small. However, in many disclosure problems of interest this will not be the case. In these circumstances, a modelling approach such as using log-linear models [19] or the independence of agreement patterns approach in section 4 seems needed. Note that to estimate $p_{M|\gamma}$ in (4) we only need to estimate the ratio $m(j)/u(j)$, which we may approximate in this case by $m(j)/u(j) = \theta_{jj} /(\tilde{f}_j / n_1)$. The factor f_j / n_2 cancels out and the key unknown required to estimate $m(j)$ is θ_{jj}. We suggest that it will normally not be realistic to expect that the intruder will be able to estimate this parameter reliably from the available data (although the mixture model approach merits further investigation). Thus, we suggest that a more realistic approach is that it is estimated by making some approximating assumptions and using external evidence on the misclassification process, as discussed above.

6 Conclusion

This risk of identification may be defined as the probability of a correct match for attacks where the intruder uses record linkage. It has been shown that expressions for this probability may be obtained for probabilistic record linkage in some special cases. In particular, expressions for the probability in the case of categorical key variables have close connections to those in other literature on disclosure risk, such as

[10]. It has also been shown that an intruder may be able to estimate these probabilities reliably under certain assumptions.

References

[1] Domingo-Ferrer, J. and Torra, V. A quantitative comparison of disclosure control methods. In: P. Doyle, J.I. Lane, J.J.M. Theeuwes, L.V. Zayatz (eds.) Confidentiality, Disclosure and Data Access. Amsterdam: North-Holland (2001)

[2] Domingo-Ferrer, J. and Torra, V. Disclosure risk assessment in statistical microdata protection via advanced record linkage. Statistics and Computing, 13, (2003) 343-354.

[3] Fienberg, S.E. Privacy and confidentiality in an e-commerce world: data mining, data warehousing, matching and disclosure limitation. Statistical Science, 21, (2006) 143-154.

[4] Spruill, N.L. Measures of confidentiality. Proc. Surv. Res. Sect. Amer. Statst. Ass (1982) 260-265

[5] Lambert, D. Measures of disclosure risk and harm. Journal of Official Statistics, 9, (1993) 313-331.

[6] Winkler, W.E. Masking and re-identification methods for public use microdata: overview and research problems. In: J. Domingo-Ferrer and V. Torra (eds.) Privacy in Statistical Databases. Lecture Notes in Computer Science 3050, Berlin: Springer, (2004) 231-246.

[7] Torra, V., Abowd, J.M. and Domingo-Ferrer, J. Using Mahalanobis distance-based record linkage for disclosure risk assessment. In: J. Domingo-Ferrer and L. Franconi (eds.) Privacy in Statistical Databases. Lecture Notes in Computer Science 4302, Berlin: Springer, (2006) 233-242

[8] Federal Committee on Statistical Methodology Statistical Policy Working Paper 22 (2nd Version): Report on Statistical Disclosure Limitation Methodology, Office of Management and Budget, Washington, D.C. (2005)

[9] Reiter, J. Estimating risks of identification disclosure in microdata. Journal of the American Statistical Association, 100, (2005) 1103-1112.

[10] Skinner, C.J. The probability of identification: applying ideas from forensic science to disclosure risk assessment. Journal of the Royal Statistical Society, Series A, 170, (2007) 195-212.

[11] Fellegi, I.P. and Sunter, A.B. A theory for record linkage. Journal of American Statistical Association, 64, (1969) 1183-1210.

[12] Bethlehem, J.G., Keller, W.J. and Pannekoek, J. Disclosure control for microdata. Journal of the American Statistical Association, 85, (1990) 38-45.

[13] Belin, T.R. and Rubin, D.B. A method for calibrating false-match rates in record linkage. Journal of American Statistical Association, 90, (1995) 694-707.

[14] Duncan, G. and Lambert, D. The risk of disclosure for microdata. Journal of Business and Economic Statistics, 7, (1989) 207-217.

[15] Larsen, M.D. and Rubin, D.B. Iterative automated record linkage using mixture models. Journal of American Statistical Association, 96, (2001) 32-41.

[16] Jaro, M.A. Probabilistic linkage of large public health data files. Statistics in Medicine, 14, (1995) 491-498.

[17] Jaro, M.A. Advances in record linkage methodology as applied to matching the 1985 Census of Tampa, Florida. Journal of American Statistical Association, 84, (1989) 414-420.

[18] Herzog, T.N., Scheuren, F.J. and Winkler, W.E. Data Quality and Record Linkage Techniques. New York: Springer (2007)

[19] Skinner, C.J. and Shlomo, N. Assessing disclosure risk in survey microdata using log-linear models. Journal of American Statistical Association, (2008) to appear.