
# Re-identification in the absence of common matching variables

Duncan Smith
duncan.g.smith@manchester.ac.uk

**Abstract**

A basic concern in statistical disclosure control is the re-identification of individuals via record linkage. A record containing identifying information in file $A$ is linked to a record containing sensitive information in file $B$, resulting in a breach of confidentiality. The classical approach to record linkage exploits the data in fields that are common to files $A$ and $B$. A more recent approach has attempted record linkage in the absence of common fields via the extraction of structural information using ordered weighted averaging (OWA) operators. Although this can be shown to perform better than a random matching strategy, it is debatable whether it demonstrates a significant disclosure risk. This paper shows that a relatively simple Bayesian approach can consistently outperform OWA linkage. Furthermore, it can demonstrate a significant risk of re-identification for the types of data release considered in the OWA record linkage literature (where there exists a 1 to 1 correspondence between the records in $A$ and the records in $B$). The Bayesian approach flows from the same underlying theory as classical record linkage, offering the possibility of using it to improve record linkage performance in more general settings.

**Keywords:** Record linkage; Re-identification; OWA operators; Bayes

**University of Manchester, CMIST Working Paper 2016-02**
http://www.cmist.manchester.ac.uk/research/publications/working-papers

# 1 Introduction

Statistical disclosure inevitably involves linkage. If sensitive information can be linked to a population member with sufficient certainty, then we have statistical disclosure. Measures to reduce the disclosure risk of released data often involve either suppression or perturbation of the raw data. A commonly used form of suppression is anonymization – the removal of directly identifying variables such as name and address. Complete records might be suppressed via sampling. Perturbation generally involves changing variable values, for instance by adding random noise to numeric variables. The success of a disclosure limitation scheme is measured in terms of both disclosure risk and data utility. The released data will hopefully be fit for purpose, and without representing a significant risk of harm to the individuals in the data or the Data Stewardship Organisation (DSO) responsible for data release.

Risk assessment often starts with an attack scenario – a description of the means by which a *data intruder* might attack a dataset (Elliot and Dale, 1999). Typically this will include a set of key variables and a set of target variables. The set of key variables is the intersection of the variables known to the intruder (regarding the population members) and the variables contained in the released data. The target variables are the sensitive variables contained in the data, other than key variables. It could be argued that if there are no sensitive variables then there is no risk of disclosure, although we must also recognize that any discovered information could potentially be used to link against additional datasets. The question addressed here is how a data intruder might attack data in the absence of key variables. We assume that the information relating to known individuals is in a file $A$, and we wish to assess the risk to a DSO of releasing a file $B$. $A$ and $B$ share no common fields.

Section 1 of the paper reviews classical record linkage. Classical linkage is used as a part of the OWA approach, and the underlying theory can be used to justify the Bayesian alternative. Section 3 describes the OWA approach, section 4 presents details of the Bayesian alternative, section 5 describes experiments that compare these approaches, with results in section 6. Risk assessment is discussed in section 7.

# 2 Record linkage

In classical record linkage (Fellegi and Sunter, 1969) we have two databases $A$ and $B$ and seek to identify record pairs that correspond to the same population units. $A$ and $B$ are assumed to be independent samples from a common

population. Fellegi and Sunter's approach is essentially a Bayesian approach, although it is not presented as such in their paper.

We have the set of all possible matches,

$$A \times B = \{(a, b); a \in A, b \in B\}.$$

This can be partitioned into sets of correctly matched and unmatched pairs,

$$M = \{(a, b) : a = b, a \in A, b \in B\}$$

$$U = \{(a, b) : a \neq b, a \in A, b \in B\}.$$

We can also partition the set of variables $X$ contained in $A$ or $B$,

$$X_A = \{x : x \in A, x \notin B\}$$

$$X_B = \{x : x \notin A, x \in B\}$$

$$X_{AB} = \{x : x \in A, x \in B\}$$

Bayes theorem leads to the following expression for the posterior odds that $a$ and $b$ correspond to the same population unit,

$$\frac{Pr((a, b) \in M|a, b)}{Pr((a, b) \in U|a, b)} = \frac{Pr(a, b|(a, b) \in M)}{Pr(a, b|(a, b) \in U)} \frac{Pr((a, b) \in M)}{Pr((a, b) \in U)} \tag{1}$$

which can be factored,

$$\begin{aligned} \frac{Pr((a,b)\in M|a,b)}{Pr((a,b)\in U|a,b)} = & \frac{Pr(\{a_v:v\in X_A\},\{b_v:v\in X_B\}|\{a_v:v\in X_{AB}\},\{b_v:v\in X_{AB}\},(a,b)\in M)}{Pr(\{a_v:v\in X_A\},\{b_v:v\in X_B\}|\{a_v:v\in X_{AB}\},\{b_v:v\in X_{AB}\},(a,b)\in U)} \\ & \times \frac{Pr(\{a_v:v\in X_{AB}\},\{b_v:v\in X_{AB}\}|(a,b)\in M)}{Pr(\{a_v:v\in X_{AB}\},\{b_v:v\in X_{AB}\}|(a,b)\in U)} \times \frac{Pr((a,b)\in M)}{Pr((a,b)\in U)} \end{aligned} \tag{2}$$

Fellegi-Sunter only considers the evidence in the key variables $X_{AB}$, and adopts the naive Bayes assumption. Values are compared only on the basis of equality leading to,

$$\frac{Pr((a, b) \in M|a, b)}{Pr((a, b) \in U|a, b)} = \left( \prod_{v \in X_{AB}} \frac{Pr(a_v = b_v|(a, b) \in M)}{Pr(a_v = b_v|(a, b) \in U)} \right) \frac{Pr((a, b) \in M)}{Pr((a, b) \in U)}. \tag{3}$$

$m$ and $u$ probabilities are defined as,

$$m_v = Pr(a_v = b_v | (a, b) \in M)$$

$$u_v = Pr(a_v = b_v | (a, b) \in U).$$

Thus the Bayes factor is a product of terms in the form $m_v/u_v$ or $(1 - m_v)/(1-u_v)$ depending upon whether the $a_v$ and $b_v$ are equal. $m$ probabilities less than one allow for distortions in the data.

The log (to the base 2) of the Bayes factor is termed a match weight in Fellegi and Sunter (1969). Thresholds on the match weights are used to allocate possible matches to one of three sets:

$A_1$ – a set of correct links
$A_2$ – a set of uncertain links
$A_3$ – a set of incorrect links

Pairs of records allocated to $A_2$ are subjected to clerical review - they are manually inspected and subsequently allocated to either $A_1$ or $A_3$. Fellegi and Sunter (1969) present a decision rule that can be used to generate thresholds corresponding to specified conditional error rates.

There are a number of approaches for estimating $m$ and $u$ probabilities and the marginal probability of a correct match, $p = Pr((a, b) \in M)$ . As the proportion of possible matches that are correct will often be very low, the $u$ probabilities can be estimated from the proportion of possible matches where variable values are equal. For certain problems the population size might be known, in which case $p$ is simply the reciprocal of the population size.

Jaro (1989) presents an Expectation Maximization (Dempster et al., 1977) algorithm for generating maximum likelihood estimates of all the required parameters.

# 3 OWA linkage

Torra (2004) describes the use of ordered weighted averaging (OWA) operators for record linkage (and therefore re-identification) when the two files $A$ and $B$ contain no common variables. The idea behind this approach is that files $A$ and $B$ will often contain common structural information, and that this can be extracted via OWA operators. Each operator is used to construct a new variable, and these variables are then used to perform classical record linkage.

## 3.1 OWA operators

An OWA operator of dimension $N$ can simply be specified as a vector $W = [w_1, \ldots, w_N]$ of non-negative weights that sum to one. The operator computes the weighted mean of an ordered vector,

$$OWA(x_1, \ldots, x_N) = \sum_{j=1}^{N} w_j y_j$$

where $y_j$ is the $j$th largest of the $x_i$.

It is possible to specify OWA operators that will calculate common summary statistics such as the minimum, maximum, mean or median.

An alternative way to specify an OWA operator is via a process that will generate the weight vector for a given $N$. Torra (2004) achieves this via a *non-decreasing fuzzy quantifier*. This is simply a non-decreasing function $F$ with domain [0,1] and range[0,1]. The weights for an OWA operator of length $N$ are then calculated as,

$$w_i = F\left(i/N\right) - F\left((i-1)/N\right).$$

Thus a non-decreasing fuzzy quantifier is a specification of an OWA operator that can be applied to vectors of different lengths.

### 3.1.1 Linkage

Initially all the variables in $A$ and $B$ are normalised by either a translation to the unit interval (range normalisation), or via standardizing so that the variable values have mean 0 and variance 1. Then a set of OWA operators are applied to the records of A and B to construct new variables. The non-decreasing fuzzy quantifier specification is used to handle differing numbers of variables in $A$ and $B$.

Two new files, $A$' and $B$', are created with numbers of rows equal to the numbers of rows in A and $B$ respectively, and with a column for each OWA operator. Each OWA operator is applied to each row of $A$ and $B$ and the resulting *representative* value placed in the relevant row and column of the relevant new file. The representatives generated by a common OWA operator are treated as the same variable for record linkage purposes.

Nin and Torra (2005) present the results of linkage experiments using data from the UCI machine learning repository (Murphy and Aha, 1994). The authors list three main assumptions:

1. The files contain a large set of correct matches

2. Both data files contain similar structural information

3. Structural information can be expressed by means of numerical representatives of individuals

There are few details regarding the record linkage approach, although they do use a classical (Fellegi-Sunter) approach. The authors demonstrate that when $A$ and $B$ tend to separate pairs of highly correlated variables (one in $A$ and one in $B$), then it is possible to achieve better linkage performance than by randomly pairing records from $A$ and $B$.

# 4 A simple Bayesian alternative

In assessing the risk of statistical disclosure we should take into account all the useful information held by a data intruder. This is not only the data that they might hold regarding individuals, but also information regarding the relationships between variables. The OWA approach attempts to exploit this information in a relatively unsupervised manner. However, we must assume that an intruder would be willing to exploit all prior knowledge or training data that were available. So here we outline a supervised learning approach that a data intruder might adopt in preference to the OWA approach.

Fellegi-Sunter linkage only exploits the data in variables common to $A$ and $B$. Without such variables we need to exploit the data ignored by Fellegi-Sunter. From equation (1) we immediately get,

$$\frac{Pr((a,b) \in M|a,b)}{Pr((a,b) \in U|a,b)} = \frac{Pr(a,b)}{Pr(a)Pr(b)} \frac{Pr((a,b) \in M)}{Pr((a,b) \in U)}. \tag{4}$$

We have two estimation problems. We need to estimate the Bayes factor, and we also need to estimate $p = Pr((a,b) \in M)$ if we want to produce posterior odds or probabilities. Firstly, we note that $p$ can be estimated from a vector of Bayes factors using Expectation Maximization, just as in Jaro (1989). For any given $p$ we can generate a vector of posterior match probabilities over the record pairs. The mean of the posterior probabilities is an estimator for $p$. So given a starting value for $p$ we can iteratively generate new posterior probabilities (expectation step) and new estimates for $p$ (maximization step). We iterate until the absolute difference between consecutive estimates of $p$ is less than some very small value.

For the Bayes factor the univariate marginals $Pr(a)$ and $Pr(b)$ could potentially be estimated from $A$ and $B$ respectively, leaving us with the problem of estimating $Pr(a,b)$ . We could also re-express the Bayes factor,

$$\frac{Pr(a,b)}{Pr(a)Pr(b)} = \frac{Pr(a|b)}{Pr(a)}$$

leaving us with the problem of estimating $Pr(a|b)$ .

Here we choose to estimate the terms in the Bayes factor in equation (4) via a full probability modelling approach exploiting the theory of decomposable graphical models.

## 4.1 Decomposable graphical models

A decomposable graph is an undirected graph $G(V, E)$ that contains no unchorded cycles of length greater than three. Each node in the graph represents a variable, and the absence of an edge $\{v, w\}$ implies that $v$ is conditionally independent of $w$ given the variables in $V \setminus \{v, w\}$. A decomposable graph can also be represented as a cluster tree. Each maximal pairwise connected subgraph of G is a cluster, and clusters are connected into a tree (or forest) so as to respect the the running intersection property (Lauritzen and Spiegelhalter, 1988):

> *If a node is contained in two clusters, $C_1$ and $C_2$, then it is contained in all clusters on the unique path between $C_1$ and $C_2$.*

Each edge in the cluster tree is associated with a sepset, the intersection of the node sets associated with the clusters that it connects. A cluster tree implies a factorization over the joint distribution of the variables in $V$,

$$Pr(V) = \frac{\prod_{C \in C} Pr(C)}{\prod_{S \in S} Pr(S)}$$

where $C$ is the set of clusters in the cluster tree (or forest) and $S$ is the set of sepsets.

For categorical variables the marginal distributions associated with clusters are marginal probability tables. The tables for sepsets can be generated by marginalisation from cluster tables. Given a structural model the table parameters can be estimated from data via maximum likelihood or via Bayesian estimation using a Hyperdirichlet prior.

Posterior beliefs over clusters given observed evidence can be generated via message passing in a cluster tree (Lauritzen and Spiegelhalter, 1988). This exploits conditional independencies and avoids calculating $Pr(V)$. Posterior beliefs over sets of variables not contained in a single cluster can be generated via variable firing (Jensen, 1996) or, at least as efficiently, by manipulating the tree so that the relevant variables appear in a single cluster (Smith, 2001).

## 4.2 Model determination

Model determination algorithms for decomposable graphical models generally depend on two important results. Frydenberg and Lauritzen (1989) showed that it is possible to move between any pair of decomposable graphs, $G$ and $G'$, by iteratively adding or removing only a single edge at a time while remaining within the class of decomposable graphs.

The basic rules for edge addition / deletion in decomposable graphs are:

> *An edge $\{v, w\}$ can be added if, and only if, it is not already present, and $v$ and $w$ are either in adjacent clusters or in distinct connected components.*
>
> *An edge $\{v, w\}$ can be deleted only if, and only if, it is present in exactly one cluster.*

Dawid and Lauritzen (1993) showed that the Bayes factor for decomposable graphical models differing by one edge can be expressed as a ratio involving only four terms, all of which can be computed locally. The marginal likelihood can be factorised,

$$p(x_V|g) = \frac{\prod_{C\in C} p(x_C)}{\prod_{S\in S} p(x_S)}.$$

It then follows from the addition / deletion rules that the ratio of marginal likelihoods for models differing by a single edge can be expressed as a ratio of products with only two terms in each product.

In addition Dawid and Lauritzen (1993) show that for categorical distributions,

$$p\left(x_A\right) = \frac{\Gamma\left(\lambda\right)}{\Gamma\left(\lambda+n\right)} \prod_i \left( \frac{\Gamma\left(\lambda_i+n_i\right)}{\Gamma\left(\lambda_i\right)} \right)$$

$$\lambda = \sum_i \lambda_i$$

$$n = \sum_i n_i$$

where $i$ indexes the cells in a table of data marginalised to the variables in $A$. The $\lambda_i$ are parameters similarly derived from a Hyperdirichlet prior.

These results have typically been exploited by Markov Chain Monte Carlo (MCMC) model determination algorithms (Madigan and York, 1995). These generate a posterior distribution over the model space. Averaging over this

distribution takes into account uncertainty in the model structure and generally provides improved predictive performance (Hoeting et al., 1999).

Madigan and Raftery (1994) use an alternative model selection strategy where they reject any models that are sufficiently poorer than the best model(s). Their *Occam's razor* strategy is based on comparisons of models differing by only a single edge. If the ratio of the posterior model probability of the smaller model to that of the larger model is below a given threshold, then the smaller model and all its submodels are rejected. A model $M_0$ is defined as a submodel of $M_1$ if all the edges in $M_0$ are also in $M_1$. Search can start from an arbitrary set of candidate models. If search starts from the complete graph, then only edge removals are considered (the down algorithm). If search starts from the model with empty edge set, then only edge additions are considered (the up algorithm). Otherwise, the down and up algorithms are run in turn to generate a set of candidate models. Finally, any candidates that are sufficiently poorer than the best model(s) are also removed. The posterior probabilities of the remaining acceptable models are normalised to sum to 1 for model averaging purposes.

# 5 Experiments

Experiments were carried out to compare the OWA approach with a Bayesian approach based on decomposable graphical models. We used four of the data sets used by Nin and Torra (2005) – the abalone, dermatology, housing and ionosphere data sets from the UCI Machine Learning Repository (Murphy and Aha, 1994). The same pre-processing steps were used - non-numeric variables were recoded using integer codes, and records with missing observations were removed.

Nin and Torra (2005) reported numbers of correct matches for 1 to 1 matching (bijection) on samples of size 30 and 100, and for three distinct sets of OWA operators. Variables were partitioned into sets $A$ and $B$ so as to induce the structural information that their approach could exploit. They showed that their approach could perform significantly better than random matching. The OWA approach used here is designed to emulate the approach in Nin and Torra (2005) as faithfully as possible while, perhaps, improving on it in certain aspects. There may be significant differences in the details of the linkage approach, and in the exploition of 1 to 1 matching. Details are contained in the following subsection.

## 5.1 OWA approach

### 5.1.1 Partitioning of variables

Nin and Torra only considered highly correlated variables and adopted a strategy of deliberately separating highly correlated variables when partitioning variables into $A$ and $B$. Variables were chosen via inspection of the correlation matrix over all the variables in the relevant dataset[1]. They used a threshold of 0.7 – variables that had no correlations with other variables above 0.7 were ignored.

Here we formalize this process. A graph is constructed with variables as nodes and pairwise correlations as edge weights. From this we generate a maximum weight spanning tree using Kruskal's algorithm (Kruskal, 1956), stopping when weights are below the threshold. The tree nodes are bi-coloured so that no pair of adjacent nodes are identically coloured (this is always possible for a tree or forest). The colouring provides us with our partitioning of variables into files $A$ and $B$.

### 5.1.2 Functions

We used the same sets of functions as Nin and Torra from which to generate corresponding vectors of weights,

$$Q_1 = \{x^\alpha : \alpha \in \{0.2, 0.4, \ldots, 2\}\}$$

$$Q_2 = \{1/(1 + e^{10(\alpha - x)}) : \alpha \in \{0, 0.1, \ldots, 0.9\}\}$$

$$Q_3 = \left\{ \begin{cases} 0 if x \leq \alpha \\ 1 if x > \alpha \end{cases} : \alpha \in \{0, 0.1, \ldots, 0.9\} \right\}.$$

Note that for each function $F$ in each set we define $F(0)=0$ and $F(1)=1$.

### 5.1.3 Linkage

Comparing representatives for equality would provide very poor linkage performance. We would expect very few (if any) matches and the vast majority (if not all) comparison vectors would be vectors of zeroes. However, it is still possible to use a standard record linkage approach if we generate binary comparison vectors by other means. There is little detail of the linkage in Nin and Torra (2005), so here we choose to employ a similarity score which

[1] Personal communication with Jordi Nin

is dichotemized to generate binary comparison vectors. A threshold of 0.95 was found to provide reasonable linkage performance.

$$\text{sim}(x, y) = max(1 - |x - y|, 0)$$

Linkage used the Expectation Maximization approach detailed in Jaro (1989). A moderate degree of Bayesian regularization (dirichlet priors and maximum a posteriori estimation) was used to avoid parameter estimates of zero. The post hoc weighting scheme contained in Winkler (1990) was also used. This tends to improve linkage performance by more fully exploiting the information in similarity scores via piecewise interpolation on match weights.

### 5.1.4 1 to 1 matching

Nin and Torra considered only subsets of data containing 30 or 100 records. Thus there was a 1 to 1 correspondence between the records in $A$ and the records in $B$. This knowledge provides important additional information that can be used to improve linkage. Firstly, we can specify $p$ – it is simply the reciprocal of the size of the subset. Using a fixed $p$ can result in improved estimation of $m$ and $u$ probabilities. Secondly, we can attempt to find the *best* 1 to 1 matching. We can construct a bipartite graph connecting each record in $A$ to each record in $B$, with edge weights equal to the posterior probabilities of a correct match. A maximum weight 1 to 1 matching can be found using the Hungarian algorithm (Kuhn, 1955). The Hungarian algorithm was compared with a greedy algorithm, where we iteratively matched the highest weight record pair ($a$, $b$) such that neither $a$ nor $b$ had previously been matched. The Hungarian algorithm generally produced better linkage performance, and those are the results presented here.

### 5.1.5 Outputs

Nin and Torra reported the numbers of correctly matched record pairs for samples of 30 and 100 records for the 3 sets of functions, Q1, Q2 and Q3. No indication of the variability in these figures across various samples was presented. For the experiments presented in the results section we provide the mean numbers of correct matches for 100 random samples.

## 5.2 Bayesian approach

Using a threshold of 0.7 for partitioning the variables resulted in a relatively low numbers of variables allocated to $A$ and $B$. Although it would have been

feasible to use MCMC or the Occam's razor approach in Madigan and Raftery (1994), we also wanted to consider other partitioning schemes and present a more generally applicable approach. Two of the chosen datasets contain 35 variables. Highly dimensional datasets are computationally problematic for the aforementioned approaches, so here we chose to use a simpler approach that searches for a single locally optimum model. We also have a preference for sparse models, not only with respect to Occam's razor, but also for the reduced computational cost of performing inference.

In common with other approaches we base our full probability modelling on adding and removing single edges while remaining within the class of decomposable graphs. We also choose to work with categorical variables – categorizing continuous variables as necessary. This reflects the fact that the majority of datasets that we will be considering in practice will contain relatively few continuously scaled variables, and those that do will often categorize these variables for disclosure risk limitation purposes. It also allows us to exploit the result of Dawid and Lauritzen (1993) presented earlier.

We use a greedy algorithm that has some similarity with the Occam's razor approach. We start with a single candidate model. In an upwards search we iteratively improve the model by adding whichever edge produces the greatest increase in posterior model probability. We stop when no improvement is possible. In a downwards search we iteratively improve the model by removing the single edge that produces the greatest increase in posterior model probability. We alternate between upwards and downwards searches until no improvement is possible.

The final model is locally optimal, but in general many local optima exist and choice of initial model is highly influential on the selected model. The goal for the present application is to find a reasonably good model in a reasonable time, while acknowledging that a sufficiently motivated data intruder might be able to do better. For computational efficiency of inference it also helps if the model is reasonably sparse. Thus we chose to start with the model with no edges (full independence model). This tends to produce much sparser models than starting with the fully connected graph (full dependence model). Experimentation showed that with a more manageable number of variables this often produced the same model as the highest posterior probability model under the Occam's razor scheme.

For the Bayesian approach the sampled observations were used for matching, while the remaining data were used for model determination. All continuous variables were split into 8 categories so that the data were evenly distributed across the categories. Model determination used Hyperdirichlet priors with all parameters equal and summing to 1. Subsequent estimation of probability tables used the same prior in order to avoid probabilities of 0.

Table 1: Numbers of variables included at various thresholds

| | 0.7 | 0.5 | -∞ |
|---|---|---|---|
| abalone | 7 | 8 | 9 |
| dermatology | 15 | 24 | 35 |
| housing | 5 | 11 | 14 |
| ionosphere | 5 | 20 | 34 |

Again, we exploited the Hungarian algorithm for 1 to 1 matching. We report results for the same random samples generated for the OWA approach.

# 6 Results

As well as the results for a threshold of 0.7 (on pairwise correlations) we also considered thresholds of 0.5 and -∞. These were used to investigate the impacts of including additional variables, and all variables respectively (one variable in the Ionosphere data set contains only a single value and was removed).

## 6.1 Tables

Tables 2 and 3 show the numbers of correctly matched records for simulations using range normalization for OWA. Standardization was not used as the differences between representative values would not have been bounded and the choice of similarity score would have been less obvious. The numbers of correct matches reported by Nin and Torra are shown in braces. The largest proportion of matches within each dataset / threshold combination are shown in bold typeface.

Table 2: Mean numbers of correct matches for random samples of size 30

| | 0.7 | | | | 0.5 | | | | -∞ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Q1 | Q2 | Q3 | Bayes | Q1 | Q2 | Q3 | Bayes | Q1 | Q2 | Q3 | Bayes |
| abalone | 11.70 | **11.93** | 10.23 | 10.28 | 7.19 | 8.22 | 7.76 | **10.07** | 3.11 | 4.63 | 5.72 | **10.97** |
| | (2) | (1) | (3) | | | | | | | | | |
| dermatology | 1.66 | 1.81 | 2.18 | **4.23** | 2.03 | 1.42 | 1.79 | **5.15** | 1.79 | 1.72 | 1.47 | **6.77** |
| | (0) | (1) | (8) | | | | | | | | | |
| housing | 3.05 | 2.42 | 1.70 | **10.23** | 1.60 | 1.60 | 1.86 | **17.17** | 1.68 | 1.18 | 1.31 | **18.97** |
| | (4) | (2) | (2) | | | | | | | | | |
| ionosphere | 5.22 | 6.05 | 5.69 | **6.24** | 7.93 | 12.16 | 15.82 | **18.47** | 4.83 | 9.09 | 13.90 | **20.38** |
| | (5) | (10) | (2) | | | | | | | | | |



Table 3: Mean numbers of correct matches for random samples of size 100

| | 0.7 | | | | 0.5 | | | | -∞ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Q1 | Q2 | Q3 | Bayes | Q1 | Q2 | Q3 | Bayes | Q1 | Q2 | Q3 | Bayes |
| abalone | **14.87** | 12.97 | 14.38 | 13.06 | 9.48 | 10.82 | 8.46 | **13.60** | 2.38 | 5.20 | 6.22 | **13.87** |
| | (3) | (7) | (2) | | | | | | | | | |
| dermatology | 1.79 | 1.71 | 2.16 | **4.36** | 2.08 | 1.31 | 1.78 | **6.05** | 2.12 | 2.50 | 1.49 | **8.18** |
| | (3) | (0) | (1) | | | | | | | | | |
| housing | 2.62 | 2.64 | 1.28 | **11.79** | 1.82 | 1.47 | 1.21 | **33.97** | 1.62 | 1.07 | 1.22 | **37.90** |
| | (1) | (2) | (1) | | | | | | | | | |
| ionosphere | 5.88 | **7.47** | 6.41 | 6.38 | 12.38 | 23.94 | 31.69 | **35.28** | 5.75 | 17.23 | 31.88 | **41.85** |
| | (12) | (21) | (3) | | | | | | | | | |

We note that there appear to be some differences between the OWA results and those reported in Nin and Torra (2005). This could be a result of the partitioning of variables, the record linkage approach, or the use of the Hungarian algorithm for 1 to 1 matching. We also note that although many possible comparisons are highly statistically significant we place little weight on this and do not report $p$-values. There are simply too many parameters that can be varied in both the OWA and Bayesian approaches that will affect performance. We restrict ourselves to the more general conclusions that we can reach from examination of the tables.

Firstly we note that the OWA approach performs relatively poorly with the Dermatology and Housing datasets. The mean numbers of correct matches are low across all thresholds. For the Abalone dataset we have a decline in performance for OWA as the threshold is reduced, whereas for the Ionosphere dataset performance is better at a threshold of 0.5.

The Bayesian approach seems to generally benefit from the inclusion of additional variables. For each dataset and sample size the best performance is achieved with the inclusion of all variables. In fact the Bayesian approach including all variables provides the largest mean number of correct matches for all datasets and sample sizes, except for the Abalone dataset where performance is similar.

Partitioning variables using Kruskal's algorithm is a device to show how effective the approaches might be in a more or less ideal situation (for the data intruder). A data intruder who simply wants to discredit a DSO might be in a position to attack a number of datasets, and might seek to find one that contains variables that have high pairwise correlations with variables known to the intruder. More typically an intruder might attack a specific dataset and have to deal with whatever variables it contains. In this situation the data intruder using the OWA approach would have to exclude some variables to optimise the attack, while the intruder using the Bayesian approach would use all variables. So the most appropriate comparison is perhaps with randomly partitioned variables where the OWA intruder removes all variables from the target dataset, $B$, that do not have at least one correlation above a given threshold with a variable in $A$. Simulation results for randomly partitioned variables are shown in Table 4. A threshold of 0.7 was used for OWA, except for the Ionosphere dataset where a threshold of 0.5 produced better performance. Again these are mean numbers of correct matches over 100 randomly generated samples. On each iteration the variables are randomly partitioned such that the maximum difference in partition size is 1.

Performance tends to be generally worse than under the original partitioning scheme. Again we find that the Bayesian approach tends to be superior to the OWA approach, except for the Abalone dataset. Fortunately

Table 4: Mean numbers of correct matches using random partitioning of variables

| | 30 | | | | 100 | | | |
|---|---|---|---|---|---|---|---|---|
| | Q1 | Q2 | Q3 | Bayes | Q1 | Q2 | Q3 | Bayes |
| abalone | 8.28 | 8.08 | 7.02 | **8.30** | **9.33** | 8.94 | 8.35 | 9.26 |
| dermatology | 2.09 | 2.31 | 2.36 | **6.17** | 2.32 | 2.39 | 2.47 | **7.04** |
| housing | 2.09 | 1.88 | 1.74 | **17.79** | 1.97 | 1.90 | 1.56 | **30.25** |
| ionosphere | 6.21 | 8.78 | 10.11 | **18.17** | 8.78 | 13.93 | 15.91 | **33.25** |

Table 5: Details of variables for the Abalone dataset

| Index | Name | Data type | Measure | Description |
|---|---|---|---|---|
| 0 | Sex | Nominal | | M, F and I (infant) |
| 1 | Length | Continuous | mm | Longest shell measurement |
| 2 | Diameter | Continuous | mm | Perpendicular to length |
| 3 | Height | Continuous | mm | With meat in shell |
| 4 | Whole weight | Continuous | grams | Whole abalone |
| 5 | Shucked weight | Continuous | grams | Weight of meat |
| 6 | Viscera weight | Continuous | grams | Gut weight (after bleeding) |
| 7 | Shell weight | Continuous | grams | After being dried |
| 8 | Rings | Integer | | +1.5 gives the age in years |

there are published metadata (Murphy and Aha, 1994) so we can investigate why the difference in performance is less marked for the Abalone data.

## 6.2 Abalone data

The Abalone data contains 9 variables. The first variable Sex is nominal, and on preprocessing has its values replaced with integer codes. All other variables are numeric, and all but the final variable (Rings) relate to Abalone size[2]. All the 'size' variables are highly correlated.

The graph in Figure 1 shows the decomposable graphical model fitted from the whole data set with nodes labelled by variable index. The light grey and dark grey nodes represent the partitioning of variables at the 0.7 threshold using Kruskal's algorithm. So all the size variables have been included. Clearly any OWA operator is going to generate some summary measure of size, and it is no surprise that these can be used for linkage purposes. This explains the similar performance of the OWA approach when

[2]An abalone is a type of edible sea mollusc

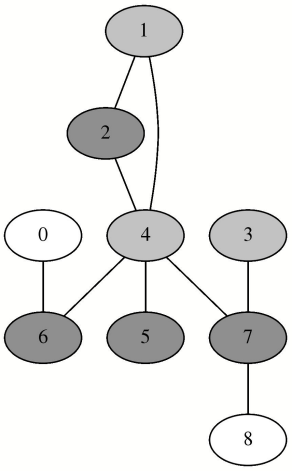


Figure 1: Decomposable graphical model for the Abalone dataset

compared with the Bayesian approach.

For the threshold of 0 the variable Rings (with index 8) is also coloured light grey. The graph suggests that Rings is conditionally independent of the other variables given Shell Weight. In fact it is approximately conditionally independent of the remaining variables given any size variable. Thus its inclusion does little more than add noise to the OWA approach. Similarly, it is likely to be relatively uninformative for the Bayesian approach. The threshold of -$\infty$ additionally includes the final variable Sex coloured dark grey. This is similarly uninformative and results in another drop in performance for OWA. This is less marked for the Bayesian approach as it less reliant on separating highly correlated variables and will not be affected by the arbitrary integer coding of categorical variables in the same way as the OWA approach. In many ways the Abalone dataset is ideal for the OWA approach, which explains why performance is comparable to the Bayesian approach, even when the Bayesian approach exploits additional variables. Given this, it is perhaps a little surprising that the results reported for the Abalone data in Nin and Torra (2005) are not better.

We can contrast the above with the Housing data set. There are few large positive pairwise correlations and of 14 variables only 5 are included in partitions at the 0.7 threshold. Performance is relatively poor across the board for the OWA approach. It does not appear to perform much better than a random matching strategy for a thresholds of 0 or -$\infty$. On the other hand, there is plenty of structure for the Bayesian approach to exploit, and as larger numbers of variables are considered performance increases substantially.

## 6.3 Precision Recall

We have seen that linkage performance drops when we use randomly partitioned variables. We might also expect it to drop if we do not have 1 to 1 matching that can be exploited by the Hungarian algorithm. That is not to say that we will not have structural information to exploit. We might have the constraint that each record in $A$ can map to at most one record in $B$ (injection) and vice versa. In some cases we might need to entertain the possibility of duplicate records within a file. We can compare the OWA and Bayesian approaches without the benefits of post-processing by generating precision-recall plots. The plots in Figures 2 and 3 were generated from the simulations used to generate Table 4. Results were aggregated over all 100 randomly generated partitions in order to assess the general performance of OWA and the Bayesian approach as classifiers.

For any given threshold on a score (here the posterior probability of a correct match) we will have a number of false positives $fp$, and a number of false negatives $fn$. Similarly we will have a number of true positives $tp$, and a number of true negatives $tn$.

$$Precision = \frac{tp}{tp + fp}$$

$$Recall = \frac{tp}{tp + fn}$$

A plot of precision against recall allows the comparison of record linkage approaches. Good approaches will produce curves in the upper right of the plot. The area under the curve is sometimes used as a performance metric.

Although these curves are based on the posterior probabilities of a correct match generated by the classifier, the curves generated on the basis of Bayes factors would be identical due to the constant marginal probability of a correct match. The expected performance of a random matching strategy is shown by a broken line.

The superiority of the Bayesian classifier is evident. In several cases the most probable match for the Bayesian approach (over the 100 partitions) is a correct match. This suggests that an intruder with a sufficiently high match probability could infer a correct match with some confidence. It does not however imply that such high probability matches are common. In fact the curves for individual partitions are highly variable. This is only to be expected - the approach relies on the existence of dependencies between the variables in $A$ and the variables in $B$, and would have no discriminatory power if the variables in $A$ were independent of the variables in $B$. This also

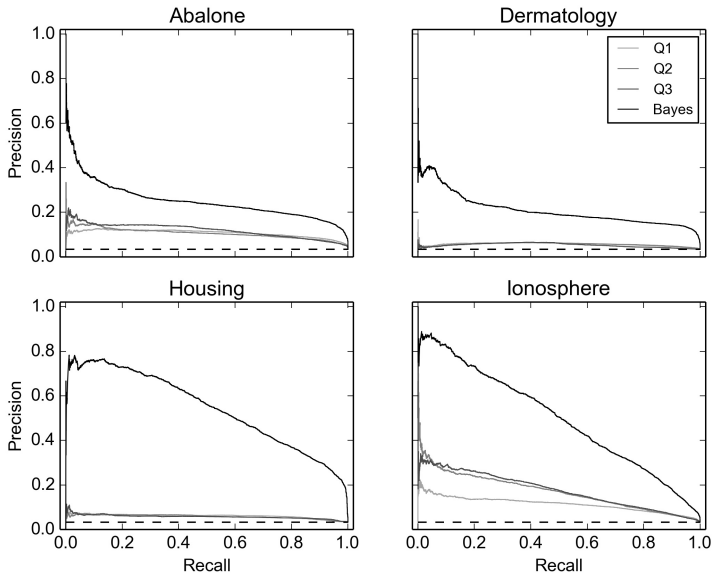


Figure 2: Precision-recall plot for $n$=30 and random partitioning (aggregated over 100 partitions)

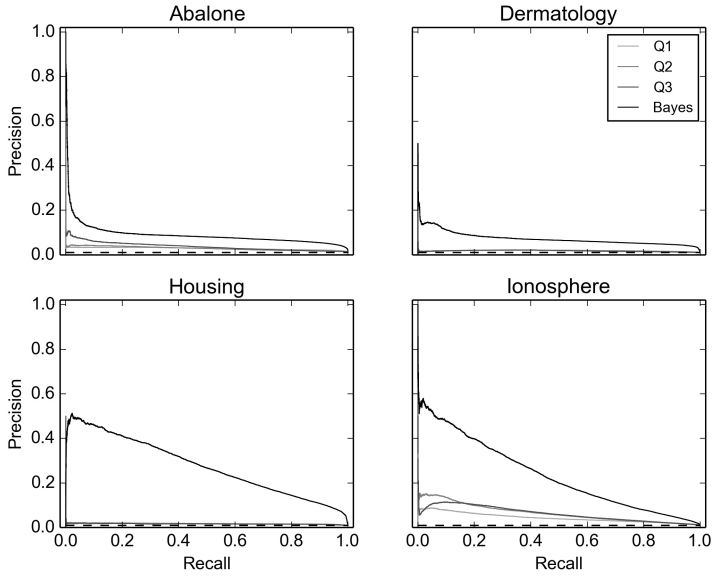


Figure 3: Precision-recall plot for $n$=100 and random partitioning (aggregated over 100 partitions)

highlights the fact that an informed intruder might be able to readily identify vulnerable datasets from prior information regarding the dependences between variables.

We also see a general decline in performance when moving from $n$=30 to $n$=100. This is to be expected due to the reduction in the marginal probability of a correct match. A priori knowledge of this marginal probability would also help in the identification of vulnerable datasets.

# 7 Risk assessment

Attack scenarios are generally paired with risk measures, and an obvious measure is the probability of a successful re-identification. The plots in Figures 2 and 3 are potentially useful for risk assessment, but the use of precision recall plots is better illustrated if we consider an alternative attack scenario.

An intruder simply seeking to discredit a DSO might attempt to maximize the probability of a successful re-identification by making a single claim of re-identification against the most probable match, and only if that match has a sufficiently high probability. So rather than constructing precision-recall plots from all possible matches we restrict consideration to those that are most probable for each randomly generated dataset / partition combination. The resulting precision-recall plots are shown in Figures 4 and 5.

The first thing to note is that the proportion of correct matches is no longer constant. Excluding all but the most probable matches from each dataset (and partition) has increased the proportion of correct matches substantially. For $n$=30 and the Abalone dataset we have proportions of 0.21, 0.52, 0.69 and 0.46 for Q1, Q2, Q3 and Bayes respectively. For $n$=30, Housing and Bayes we have a proportion of 0.83. These represent the empirical probabilities of a successful re-identification at a threshold of 0 - the proportions of most probable matches that are correct matches.

Another notable feature is that choosing higher thresholds does not consistently increase the probability of re-identification. The data intruder certainly benefits from only considering the most probable match from each dataset, but only seems to clearly benefit from further exclusion of less probable matches for the Abalone dataset with the Bayesian approach.

The results for $n$=100 are consistent with those for $n$=30. Only for the Abalone dataset and the Bayesian approach does the use of a non-zero threshold clearly increase the probability of re-identification. Risks are generally lower, although we still have a probability of successful re-identification of 0.57 for the Housing dataset and the Bayesian approach.

In practice a DSO might be interested in assessing risk for a single dataset

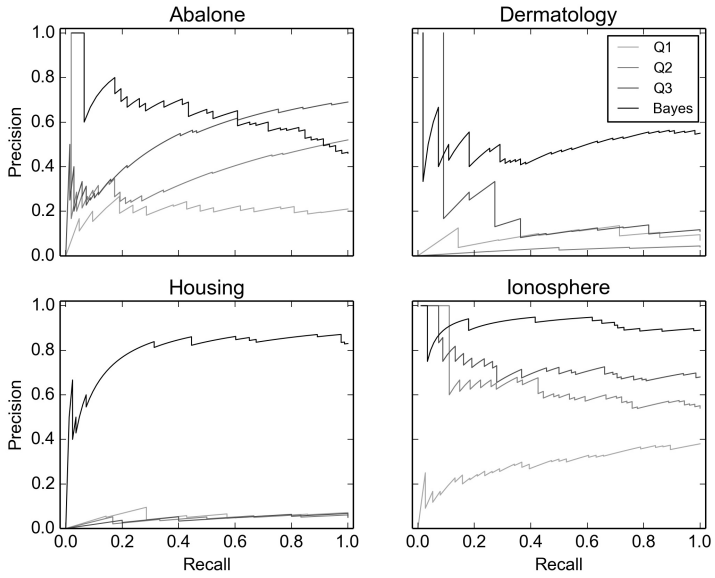


Figure 4: Precision-recall plot for $n$=30 and the most probable match strategy

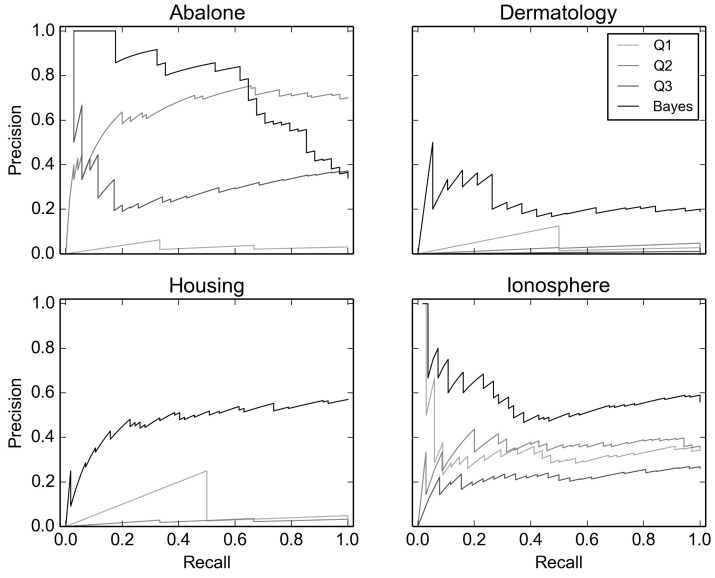


Figure 5: Precision-recall plot for $n$=100 and the most probable match strategy

or a collection of datasets. The relevant variables in $B$ might be fixed, or variable if attribute suppression is a possibility. The variables in $A$ will depend on the attack scenario under consideration. Nevertheless, as long as there is a notion of an intruder-specified threshold on the match probability, then precision-recall plots can be useful risk assessment tools. Of course, this also applies in situations where there are also common variables that can be used for matching.

# 8 Conclusions

Nin and Torra (2005) showed that the OWA approach could perform significantly better than a random matching strategy. We have shown that a relatively simple Bayesian approach can consistently outperform OWA, the exception being a rather degenerate dataset that is ideally suited to OWA. We have shown that the risks are of practical significance for the 1 to 1 matching problems considered in the OWA literature. In this case matching can be significantly improved by exploiting structural information via the Hungarian algorithm. The precision-recall analysis demonstrated that there can still be appreciable risks of re-identification when this structural information is either not exploited, or not present.

Analysis has been restricted to datasets that are considered in Nina and Torra (2005) - for the purposes of comparison. These are suited to the OWA approach as they contain large numbers of numeric variables. In practical circumstances we will tend to meet datasets containing categorical variables - not least because numeric variables are often categorised for statistical disclosure risk limitation purposes. The Bayesian approach was designed to deal with the more usual case, and numeric variables had to be categorized. We expended some effort trying to optimise the OWA approach, and hardly any effort trying to optimise the Bayesian approach. Given the above, and the difference in performance, we would have to recommend the Bayesian approach over the OWA approach for risk assessment. Precision-recall plots are useful tools for risk assessment, and can be generated for both collections of datasets and individual datasets, and under various attack scenarios.

An obvious extension to the Bayesian approach is to exploit the information in non-overlapping variables to improve classical record linkage. Equation (2) shows exactly how this can be approached. This is an area for future work. Some early work has shown that naively combining the two approaches is ineffective. We can decompose the Bayes factor into a product of terms relating to the non-key variables and key variables as presented earlier. But it seems to be important that the former term is conditioned on the key

variables. Another consideration is that Fellegi-Sunter is designed to accommodate errors (perhaps introduced via deliberate perturbation) through the $m$-probabilities. To some degree this will also be true of the OWA approach. The simple Bayesian approach presented here does not accommodate errors in Files $A$ or $B$ unless they are present in (or introduced to) the training data.

# Acknowledgements

This work has been partially supported by the National Centre for Research Methods.

# References

[1] Dawid A.P. and Lauritzen S.L. (1993) Hyper Markov laws in the statistical analysis of decomposable graphical models. *The Annals of Statistics*, 21, pp.1272-1317

[2] Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977) Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, Series B 39 (1) pp.1-38

[3] Elliot, M.J. and Dale, A. (1999) Scenarios of attack: the data intruder's perspective on statistical disclosure risk. *Netherlands Official Stat.* 14, pp.6-10

[4] Fellegi, I.P. and Sunter A.B. (1969) A theory for record linkage. *JASA* Vol. 64, No. 238, pp.1183-1210

[5] Frydenberg, M. and Lauritzen, S.L. (1989) Decomposition of maximum likelihood in mixed graphical interaction models. *Biometrika*, **76**, 3, pp.539-55

[6] Hoeting, J.A., Madigan, D., Raftery, A.E. and Volinsky, C.T. (1999) Bayesian model averaging: a tutorial. *Statistical Science*, 14(4), pp.382-417

[7] Jaro M.A. (1989) Advances in record linkage methodology as applied to the 1985 census of Tampa Florida. *JASA* 84 (406) pp.414-20

[8] Jensen, F.V. (1996) *An Introduction to Bayesian Networks.* Springer-Verlag, New York

[9] Kruskal, J.B. (1956) On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical Society* 7 pp.48-50

[10] Kuhn, H.W. (1955) The Hungarian Method for the assignment problem. *Naval Research Logistics Quarterly*, **2**: pp.83-97

[11] Madigan, D. and Raftery, A.E. (1994) Model selection and accounting for model uncertainty in graphical models using Occam's window. *JASA*, 89, pp.1535-1546

[12] Madigan, D. and York, J. (1995) Bayesian graphical models for discrete data. *International Statistical Review*, 63 pp.215-232

[13] Murphy, P.M. and Aha, D.W. (1994) *UCI repository machine learning databases*. http://www.ics.uci.edu/mlearn/MLRepository.html

[14] Lauritzen, S.L. and Spiegelhalter, D.J. (1988) Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society B*, 50, No. 2., pp.157-224

[15] Nin, J. and Torra, V. (2005) Towards the use of OWA operators for record linkage. *EUSFLAT-LFA*

[16] Smith, D. (2005) The efficient propagation of arbitrary subsets of beliefs in discrete-valued Bayesian belief networks. In *Proceedings of the Eighth International Workshop on Artificial Intelligence and Statistics* (eds T. Jaakkola and T. Richardson), Key West, Florida, pp.292-297

[17] Torra, V. (2004) OWA operators in Data Modeling and Reidentification. *IEEE Transactions on Fuzzy Systems*, Vol. 12, No. 5, pp.652-660

[18] Winkler, W. E. (1990) String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. *Proceedings of the Section on Survey Research Methods (American Statistical Association)* pp.354-359