A unified approach to the assessment of both identification and attribution risks

Duncan Smith[*], Mark Elliot[**]

[*] University of Manchester, Oxford Road, Manchester, England,
duncan.g.smith@manchester.ac.uk

[**] University of Manchester, Oxford Road, Manchester, England,
mark.elliot@manchester.ac.uk

Abstract. In statistical disclosure control we are generally concerned with either identification risk, or attribution risk. The assessment of these forms of risk depends on the type of data. When we have two databases A and B which contain common variables classical record linkage provides an appropriate mechanism for risk assessment. Measures for identification risk might be based on match probabilities. Measures for attribution risk might be based on models generated from match probabilities. In order to adequately assess risk we must endeavour to perform linkage at least as well as any potential data intruder. Here we show that the classical record linkage approach will tend to under-estimate the risk from a knowledgeable intruder. We present an improvement over classical linkage that simultaneously generates match probabilities and an estimate of the joint distribution over the variables in A and B. Thus it generates the basic information that we need to assess both identification and attribution risk.

1 Introduction

Consider a Data Stewardship Organisation (DSO) which is considering whether to release data for research, or other purposes. The question is whether the benefit of data release outweighs the risk of disclosing sensitive information. The data in question will often have had identifying information such as name and address removed. Nevertheless, there can remain some risk that sensitive information could be leaked to a *data intruder* through linkage to other data sources available to the intruder. To limit disclosure risk it is necessary to first assess the risk and, perhaps, perturb or partially suppress the data before they are released. Although the specific means of linkage will depend on the type of data, it is clear that one of the tools that could be used by the data intruder is record linkage.

1.1 Statistical disclosure risk

Statistical disclosure risk is usually described in terms of a data intruder attempting to learn sensitive information about some *target* under some *attack scenario* [Elliot and Dale, 1999]. The target is often an individual person, although it could also be an organisation. The attack scenario describes the information available to the data intruder and the strategy employed by the intruder to recover information. We are generally interested in two distinct forms of disclosure.

> Identification – a data intruder associates a published record with a target.
>
> Attribution – a data intruder associates the value of a sensitive attribute with a target.

Identification might not involve disclosure – the data intruder might not learn anything new or sensitive regarding the target. Yet it can still a concern because it suggests that published data are vulnerable to attack.

1.2 Record linkage

Record linkage is the practice of identifying pairs of records in distinct databases that relate to the same entity. The classical linkage framework [Fellegi and Sunter, 1969] addresses the linking of records between two datasets A and B that are assumed to be random samples from a common population.

There is a set of all possible matches,

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

which can be partitioned into sets of matched and non-matched pairs,

$$M = \{(a, b); a = b, a \in A, b \in B\} \qquad U = \{(a, b); a \neq b, a \in A, b \in B\}$$

and the goal of record linkage is to allocate the possible matches to these sets.

Assume the data are aligned so that each index $i \in \{1, \ldots, n\}$ corresponds to the same variable in A or B. Then, Fellegi-Sunter assumes that the posterior odds of a match can be factorized as,

$$\frac{Pr((a, b) \in M|(a, b))}{Pr((a, b) \in U|(a, b))} = \left(\prod_{i=1}^{n} \frac{Pr((a_i, b_i)|(a, b) \in M)}{Pr((a_i, b_i)|(a, b) \in U)} \right) \frac{Pr((a, b) \in M)}{Pr((a, b) \in U)}. \tag{1}$$

Variables that are not common to both A and B are ignored, comparisons of values are only for equality / inequality, and it is assumed that comparisons on a given key variable are independent of the comparisons on other key variables given the match status.

Linkage can compromise privacy in a number of ways:

1. If the records in, say, A contain identifying information, then record linkage can be used to re-identify the records in B.

2. Even if matches cannot be inferred with high certainty, some values might be attributed with high probability if they are shared by several uncertain matches to the same record.

Thus linkage can be used to assess both identification and attribution risk.

To adequately assess risk we must perform linkage to a standard comparable with any data intruder. To this end we might employ several extensions of the classical approach. String similarities can be used to account for typographical errors [e.g. Winkler, 1990, Smith and Shlomo, 2014]. Matching constraints can be incorporated [e.g. Sadinle, 2016]. The estimation of the Fellegi-Sunter parameters might be improved by exploiting labelled training data; Larsen and Rubin [2001] iteratively re-estimated parameters after manually classifying uncertain links (e.g. those with match probabilities close to 0.5) and adding them back as training data. Here we consider only one specific extension to the Fellegi-Sunter approach – the simultaneous estimation of Fellegi-Sunter parameters and a full probability model [Smith and Elliot, 2017]. Parameters are iteratively re-estimated within an expectation-maximization (EM) framework.

2 Extended linkage approach

Expectation-maximization [Dempster et al., 1977] is an iterative approach to maximum likelihood estimation. On the E-step of the algorithm 'missing' parameters in the likelihood function are replaced with their expectations given current estimates of the 'non-missing' parameters. (The algorithm is initialised with plausible estimates for non-missing parameters.) On the M-step of the algorithm we find the maximum likelihood estimates of the non-missing parameters conditional on the expectations of the missing parameters. Iterating E and M steps until convergence will find a (possibly not global) maximum of the likelihood function. Jaro [1989] presented an EM approach for Fellegi-Sunter linkage, where the unobserved match statuses were treated as missing.

In order to simultaneously estimate match probabilities and a full probability model we adopt an extended latent model [Smith and Elliot, 2017],

$$\begin{aligned}\frac{Pr((a,b)\in M|(a,b))}{Pr((a,b)\in U|(a,b))} = {} & \frac{Pr((a,b)|(a,b)\in M)}{Pr((a,b)|(a,b)\in U)}\times \\ & \left(\prod_{i=1}^{n}\frac{Pr((a_i,b_i)|(a,b)\in M)}{Pr((a_i,b_i)|(a,b)\in U)}\right)\frac{Pr((a,b)\in M)}{Pr((a,b)\in U)}\end{aligned} \tag{2}$$

where the additional Bayes factor concerns the values of variables rather than the binary comparison of value pairs. As in Jaro [1989] we treat the unobserved match statuses as missing data.

Let $\gamma_i^j = 0$ if attribute i differs for record pair j, and $\gamma_i^j = 1$ if attribute i matches for record pair j.

$$m_i = Pr(\gamma_i^j = 1 | r_j \in M) \qquad u_i = Pr(\gamma_i^j = 1 | r_j \in U) \qquad p = \frac{|M|}{|M \cup U|}$$

$$Pr(\gamma^j | M) = \prod_{i=1}^{n} m_i^{\gamma_i^j} (1 - m_i)^{1-\gamma_i^j} \qquad Pr(\gamma^j | U) = \prod_{i=1}^{n} u_i^{\gamma_i^j} (1 - u_i)^{1-\gamma_i^j}$$

We want to estimate both the unknown Fellegi-Sunter parameters $\Phi = (m, u, p)$ and parameter vectors Φ_m and Φ_u associated with the models that generate the numerator and denominator terms of the additional Bayes factor.

Let x be the complete data vector equal to $\langle \gamma, g \rangle$, where $g_j = (1{,}0)$ iff the jth record pair $r_j \in M$ and $g_j = (0{,}1)$ iff $r_j \in U$. Let the configuration of the evidence on the values of the variables for the jth record pair be denoted δ^j. Then we can show that the log likelihood for the complete data is,

$$\begin{aligned}\ln(f(x|\Phi, \Phi_m, \Phi_u)) = &\sum_{j=1}^{N} g_j \cdot \left(\sum_{i=1}^{n} \ln(m_i^{\gamma_i^j}(1 - m_i)^{1-\gamma_i^j}), \sum_{i=1}^{n} \ln(u_i^{\gamma_i^j}(1 - u_i)^{1-\gamma_i^j}) \right)^T + \\ &\sum_{j=1}^{N} g_j \cdot (\ln(p), \ln(1 - p))^T + \\ &\sum_{j=1}^{N} g_j \cdot \big(\ln(Pr(\delta^j | \Phi_m)), \ln(Pr(\delta^j | \Phi_u))\big)^T \end{aligned} \tag{3}$$

If we follow Smith [2016] we can generate the additional Bayes factors from a single full probability model,

$$Pr((a, b)|(a, b) \in M) = Pr(\{a_v : v \in A - B\} \cup \{b_v : v \in B - A\} \cup \{a_v : v \in A \cap B, a_v = b_v\})$$

$$Pr((a, b)|(a, b) \in U) = \frac{Pr(\{a_v : v \in A - B\}) Pr(\{b_v : v \in B - A\})}{Pr(\{a_v : v \in A \cap B, a_v = b_v\})}$$

in which case $\Phi_m = \Phi_u$.

Once the g_j in Formula 3 are replaced by their expectations on the E-step, then we can generate maximum likelihood parameter estimates for Φ very easily and relatively efficiently [see Jaro, 1989]. The difficulty is in generating new estimates for the Φ_m (and the Φ_u in the two model case). It is very easy to over-fit the model(s), and strategies to avoid over-fitting are the subject of ongoing research. That research will not be discussed here. Here we illustrate the approach with the following simple, but effective strategy.

We estimate the u_i from the record pairs under the assumption that all pairs are incorrect matches [see Jaro, 1989]. We perform a standard Fellegi-Sunter linkage run with fixed u_i to estimate the m_i and p. The match probabilities are then aggregated to generate a table of pseudocounts (all variables are categorical). Missing values (non-matches on key variables) are dealt with via the EM approach of Fuchs [1982]. A single decomposable graphical model is fitted from the pseudocounts using a greedy approach [Smith, 2016]. We then perform a run with the extended latent model with fixed u_i, fixed $\Phi_m = \Phi_u$, and with the m_i and p estimated from the Fellegi-Sunter run as starting values.

Decomposable graphical modelling will not be described here, Smith [2016] contains relevant details and references. We simply note that a decomposable graphical model is based on an undirected graph $G(V, E)$ with variables as nodes and properties described in e.g. Lauritzen and Spiegelhalter [1988]. The graph G admits a factorization over the joint distribution of V,

$$Pr(V) = \frac{\prod_{C \in \boldsymbol{C}} Pr(C)}{\prod_{S \in \boldsymbol{S}} Pr(S)} \tag{4}$$

where the set $\boldsymbol{C}$ and multi-set $\boldsymbol{S}$ both contain variable sets that are pairwise connected in G.

3 Disclosure control

In this section we will present some linkage results and demonstrate how they can be used for risk assessment. The data analysed were sampled (n=1000) from an existing full probability model – the insurance (Bayesian) network[1]. Several nodes were pruned[2] leaving a reduced network with nodes:

> SeniorTrain, RiskAversion, VehicleYear, GoodStudent, DrivingSkill, SocioEcon, DrivQuality, HomeBase, Age, MakeModel, OtherCar

The sample was then augmented by adding an additional variable 'Surname' which was randomly generated from an online database of U.S. surnames. The

[1] `http://www.bnlearn.com/bnrepository/#insurance` – visited 12th Nov. 2016

[2] Pruning is the (recursive) removal of leaf nodes and the associated probability distributions

sample was then sub-sampled to produce datasets A (n=400) and B (n=600) with variables:

A - SeniorTrain, RiskAversion, VehicleYear, GoodStudent, DrivingSkill, SocioEcon, DrivQuality, Surname

B – DrivQuality, Surname, HomeBase, Age, MakeModel, OtherCar

The key variables in B were randomly perturbed to simulate both typographical errors and the entry of incorrect values.

Possible matches are generally linked (classified as matches) if, and only if, the associated match probability is above a given threshold. Thus for any given threshold 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*.

$$Precision = \frac{tp}{tp + fp} \qquad\qquad Recall = \frac{tp}{tp + fn}$$

A plot of precision against recall for a large range of thresholds (one threshold for each distinct precision, recall pair) allows the comparison of record linkage approaches. The area under the curve (AUC) can be used as a summary measure of performance. Here we plot expected precision against expected recall in order to remove the noise that we would have if we simply ordered matches / non-matches arbitrarily within an equivalence class (set of record pairs associated with a given match probability).

We show results for 4 different linkage strategies:

FS – Standard Fellegi-Sunter

Ext. EM – Extended EM with no training data

Ext. EM + – Extended EM with 500 training data

GP – Fixed Bayes factors generated from the known underlying data generating process

The training data were randomly sampled from the reduced network. We could think of the training data as representing prior knowledge held by the data intruder. The GP series represents a notional gold standard – the potential for improvement that we have under the extended latent model.

Figure 1 shows the precision recall curves for the strategies listed above. We have significantly improved linkage performance for the approaches that use the extended latent model. The use of additional training data further improves performance, and the best performance is when using Bayes factors derived from the known generating process.

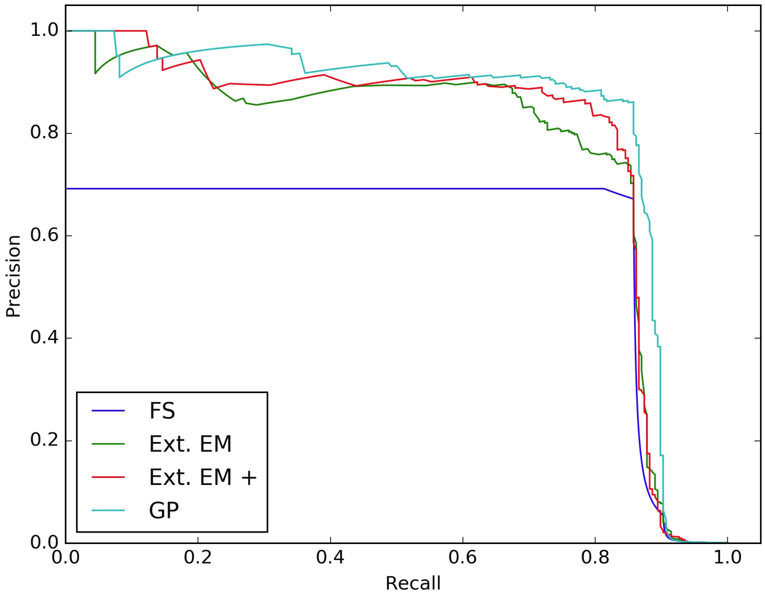

Figure 1: Precision recall curves for Fellegi-Sunter linkage, extended EM, extended EM with training data, and extended EM with (additional) Bayes factors derived from the data generating process

Precision recall curves depend only on the ordering of record pairs and their match statuses. They tell us nothing about the accuracy of the estimated match probabilities – other than how well they order the record pairs. To investigate the accuracy of the match probabilities we produced Figure 2 which plots cumulative match probabilities (ordered from largest to smallest) against the corresponding cumulative expected matches. (Again we used expected matches to remove noise due to arbitrary orderings within an equivalence class.)

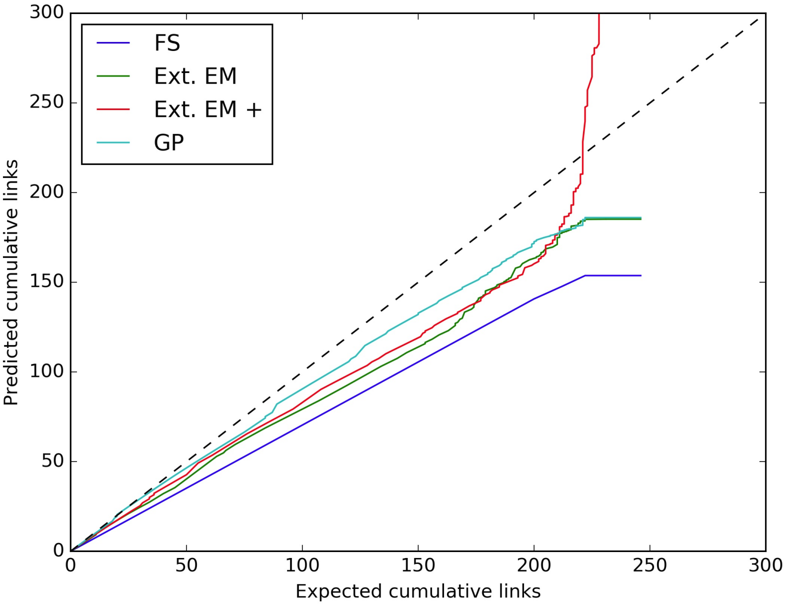

Figure 2: Plot of cumulative match probabilities against cumulative expected matches – record pairs ordered from most probable match to least probable match

It is notable that (for this linkage exercise) Fellegi-Sunter has consistently underestimated the true match probabilities. Despite 69.2% of the most probable matches being matches Fellegi-Sunter provides a match probability of only 0.487. The approaches based on the extended latent model seem to provide more accurate match probabilities. All linkage strategies assign very low match probabilities to a small proportion of matches. The extended EM with training data also assigns inflated (but still low) probabilities to a number of non-matches. This kind of behaviour is inevitable given only two key variables and the number and types of errors that we have introduced. Some matches will have no matching values on the key variables, and some non-matches will match on one or both key variables. The important point is that the extended approaches provide better estimated match probabilities for the most probable matches.

3.1 Identification risk

The actual risk depends on the detail of the strategy used by a data intruder. A DSO in possession of know match statuses can simply attack the data and calculate the probability that the attack is successful. We can see immediately from Figure 1 what the risk is for an intruder who selects a most probable match and declares it to be a re-identification. The attempted re-identification is successful with probability 0.692 under Fellegi-Sunter linkage, and with probability 1 under the extended EM strategies. Faced with this a DSO could use perturbation or suppression (of records or variables) to reduce risk, re-running the linkage to assess risk reduction. This would not necessarily involve perturbing or suppressing the riskiest record pairs. In fact it could involve the introduction of dummy records designed to negatively impact linkage performance. It is the effectiveness of the linkage exercise (with respect to the detailed attack scenario) that needs to be altered.

Without known match statuses the DSO must assess risk on the basis of the match probabilities. For the linkage exercises presented here Fellegi-Sunter has consistently under-estimated the true match probabilities. This might be considered to be a good thing for disclosure control as an intruder would be less likely to make a claim of re-identification. On the other hand, a DSO without access to known match statuses would under-estimate the risk of re-identification under the 'claim the most likely match as a match' scenario. Happily the approaches based on the extended latent model seem to provide less biased match probabilities, particular for the higher probability matches that are more likely to result in claims of re-identification.

Risk assessment should consider everything that an intruder might have in his / her armoury. Exploiting string similarities via the method of Smith and Shlomo [2014] can easily be incorporated into the extended EM approach. It is less obvious how matching constraints could be incorporated (except as a post hoc procedure). Linkage can also be improved through prior information regarding the Fellegi-Sunter parameters or the population distribution. These can all be incorporated into the extended EM approach.

3.2 Attribution risk

An intruder using the Fellegi-Sunter approach could estimate posterior probabilities via aggregation over match probabilities. Each record $a \in A$ will have a match probability to each record in B. If the higher match probabilities all correspond to members of B that share a common attribute value, then this value might be associated with a. In order to attack the data an intruder might initially construct a joint distribution over all the variables using the data and match probabilities. Missing data can be dealt with via the approach of Fuchs [1982] to generate a maximum likelihood joint distribution. Having constructed a joint distribution it is possible for the intruder to generate arbitrary conditional distributions to attack the data.

Again, the risks will depend on the detailed attack scenario, and the DSO can assess risk by attacking the data. But using the extended latent model we estimate the joint distribution as part of the linkage exercise. We do not consider detailed attack scenarios here, but propose that risks will generally be greater if the joint distribution generated by the intruder is closer to that of the population. As we generated the data from a known generating process we can generate joint distributions and compare with the known population distribution. This we do by calculating the Jensen-Shannon (JS) divergences shown in Table 1.

The 'full dependence' results are based on the use of Fuchs' algorithm (which implicitly assumes a full dependence model). The 'decomposable' results use an additional step. To generate parameter estimates for the decomposable models we simply marginalise the joint distribution above to the clusters and separators of the decomposable model. The joint distribution is then given by Formula 4. This significantly reduces the JS divergences, especially for the generating process (where we use the model that generated the data).

	Fellegi-Sunter	Ext. EM	Ext. EM +	GP
Full dependence	0.573	0.546	0.328	0.508
Decomposable	N/A	0.333	0.199	0.113

Table 1: Jensen-Shannon divergences for tables using alternative linkage strategies and model structures

4 Discussion

We have shown that record linkage performance can be improved by extending the usual Fellegi-Sunter latent model to include a Bayes factor that is the ratio of the conditional probabilities of a record pair under M and U. Not only does it improve linkage by generating better orderings of record pairs, it also generates a joint distribution over the variables. By handling missing values and respecting conditional independences (that are consistent with the match probabilities) we generate better estimates of the joint distribution than when simply aggregating match probabilities.

A DSO assuming a standard Fellegi-Sunter linkage attack will tend to underestimate both identification and attribution risk for an intruder who is using the extended latent model. DSOs should be aware of improvements to the classical linkage approach and use those improved methods when assessing risk from a knowledgeable intruder. We have noted that the extended latent model can be combined with other approaches (such as the use of similarity scores) to achieve yet greater improvements in linkage performance.

We have only presented the results of a very basic strategy under the extended latent model. We only performed modelling under the initial Fellegi-Sunter match probabilities. The extended approach allows for iterative re-estimation of the model(s) and associated Bayes factors. This can be very effective, but there are issues regarding how best do achieve this without stepping outside the EM framework and while avoiding over-fitting. This is the subject of ongoing research.

We have only addressed linkage in terms of identity – i.e. identifying record pairs that refer to the same individual or organisation. In disclosure control we might also be interested in other types of relationship, such as familial, social network or ownership relationships. This more general problem is termed *graph linkage*. We propose that the idea of simultaneously estimating the match probabilities and the parameters of a population model could also be applied to the more general problem of graph linkage. This is also the subject of ongoing research.

References

A. P. Dempster, N. Laird, and D. B. Rubin (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38.

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

I. Fellegi and A. Sunter (1969) A theory for record linkage. *JASA*, 64(238):1183–1210.

C. Fuchs (1982) Maximum likelihood estimation and model selection in contingency tables with missing data. *JASA*, 77(378):270–278.

M. A. Jaro (1989) Advances in record linkage methodology as applied to the 1985 census of Tampa Florida. *JASA*, 84(406):414–420.

M. Larsen and D. Rubin (2001) Iterative automated record linkage using mixture models. *JASA*, 96(453):32–41.

S. Lauritzen and D. Spiegelhalter (1988) Local computations with probabilities on graphical structures and their application to expert systems. *JRSS series B*, 50 (2):157–224.

M. Sadinle (2016) Bayesian estimation of bipartite matchings for record linkage. *JASA*, pages 1–35. doi: 10.1080/01621459.2016.1148612. URL `http://dx.doi.org/10.1080/01621459.2016.1148612`.

D. Smith (2016) Re-identification in the absence of common matching variables. Technical report, University of Manchester. URL `http://hummedia.manchester.ac.uk/institutes/cmist/archive-publications/working-papers/2016/2016-02.pdf`.

D. Smith and M. Elliot (2017) Improving record linkage via the application of occam's razor. *Under review*.

D. Smith and N. Shlomo (2014) Report for the data without boundaries project. Technical report, University of Manchester. URL `http://hummedia.manchester.ac.uk/institutes/cmist/archive-publications/reports/2014-01-Data_without_Boundaries_Report.pdf`.

W. E. Winkler (1990) String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage. In *Proceedings of the Section on Survey Research Methods (American Statistical Association)*, pages 354–359.