An Experiment in naïve Bayesian Record Linkage

Duncan Smith, Mark Elliot
University Of Manchester, Cathie Marsh Centre for Census and Surveys Research
Oxford Road, Manchester, M13 9PL, UK
duncan.g.smith@manchester.ac.uk, mark.elliot@manchester.ac.uk

Abstract

Sharing data can represent a risk of disclosing sensitive information about the individuals which the data sets concern. Computationally complex techniques can be used by a so-called 'data intruder' to link such data and discover information about targeted individuals. Heuristic approaches to limiting this risk are aimed towards the more casual intruder. A knowledgeable intruder, armed with data mining tools, can uncover sensitive information from ostensibly safe data sets. This paper considers a method for assessing the risk of disclosure by a relatively knowledgeable intruder, whilst avoiding the computational problems associated with exact probability calculations.

Introduction

This paper addresses the degree to which the ability of a data intruder to correctly link records between distinct sample microdata sets is affected by the co-presence of a set of publicly available aggregate population tables. It is based on an opportunistic attack scenario where an intruder has identified records in two separate samples (relating to the same population) that match on the set of variables that are common to the two sample tables. Thus it is possible that the two partial records correspond to the same individual. Specifically we consider the case where a pair of potential matching records are each unique on a set of common attributes.
The intruder is assumed to be mathematically literate and have access to reasonably standard computational resources. A relatively naïve Bayesian approach is used to assess the degrees of confidence an intruder might reasonably have in the correctness of a potential match. It is assumed that the intruder has no prior information regarding dependencies between variables.

A general inferential approach

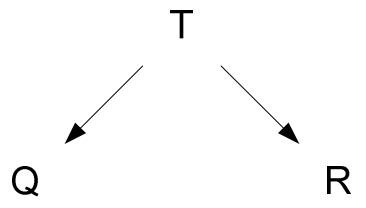

Figure.1. General causal structure

The above directed acyclic graph (DAG) represents a general causal structure relating to tabular data (T), a data release (R) and (the answer to) a relatively arbitrary query (Q). The implied assumptions are:

1. The query can be fully answered by reference to a known T.
2. The mechanism for transforming T to R is known.

In general T is not observed, R is observed and we require a posterior distribution over Q. The posterior distribution is given by,

Eqn.1. $$P(Q|R=r)=\frac{\sum_T P(Q|T)P(R=r|T)P(T)}{\sum_T P(R=r|T)P(T)}.$$

Evaluating this expression is computationally expensive because of the requirement to sum over a potentially enormous space of tables. In principle this space is infinitely large, but in practice it is possible to iterate over only those tables for which $P(R=r|T) > 0$, although this often still represents an enormous computational burden.

Rounded tables

Consider a query relating to the value of a single cell with index, say, 1 in a table, T. Assume that a perturbed version of T is released, such that each population cell count F_i is perturbed independently of the other population cell counts to give a perturbed count f_i. This is the case for many perturbation schemes, such as random rounding and deterministic rounding. Then we have the causal structure shown in Figure 2.

In general, all the F_i are children of T, and each F_i has a child (a corresponding perturbed value) for each perturbed table $r \in R$.

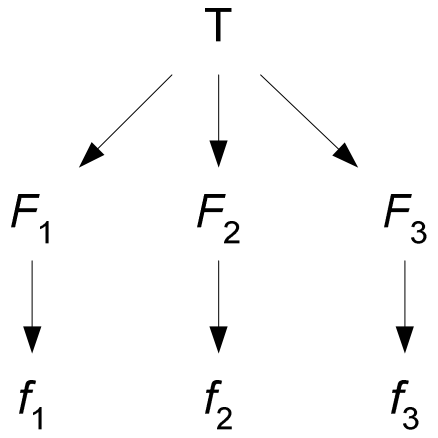

Fig.2. Cell level causal structure

Here all the f_i are observed and F_1 is queried. The posterior distribution over F_1 is given by the expression in Eqn.2.

Eqn.2.

$$p(F_1|f_1,f_2,f_3)=\frac{p(f_1|F_1)\sum_{F_2}p(f_2|F_2)\sum_{F_3}p(f_3|F_3)\sum_{T}p(F_1|T)p(F_2|T)p(F_3|T)p(T)}{\sum_{F_1}p(f_1|F_1)\sum_{F_2}p(f_2|F_2)\sum_{F_3}p(f_3|F_3)\sum_{T}p(F_1|T)p(F_2|T)p(F_3|T)p(T)}.$$

The important point is the summation over all feasible tables. However, assume an improper uniform prior over T. This implies complete independence of the F_i and inference is simplified enormously. Effectively, T can be removed from the causal graph and we (or an intruder) can avoid the summation over T. The uniform prior over T implies a uniform prior over each F_i and the posteriors for the F_i are given by,

Eqn.3.
$$p(F_i|f_i)=\frac{p(f_i|F_i)}{\sum_{F_i}p(f_i|F_i)},$$

which only requires a summation over the feasible values of F_i. The $p(f_i|F_i)$ are easily calculated from the known rounding scheme. A similar approach was used in Smith and Elliot (2003).

Sample tables

If a sample table (with known sampling fraction, implying the exact total can be accurately estimated) is released, then the likelihood $p(f_i|F_i)$, where f_i denotes the number sampled from cell F_i, is simply the Hypergeometric probability,

Eqn.4. $$p(f_i|F_i) = \frac{\binom{F_i}{f_i}\binom{N-F_i}{n-f_i}}{\binom{N}{n}}$$

where N and n denote the (estimated) population and samples sizes respectively.

Thus for a release R consisting of rounded and sampled versions of T the posterior is given by,

Eqn.5. $$p(F_i|R) = \frac{\prod_R p(f_i|F_i)}{\sum_{F_i}\prod_R p(f_i|F_i)}.$$

Other tables

In practice an intruder will not generally be lucky enough to have a set of released tables, each containing a perturbed version of a cell of interest. In some cases perturbed counts will correspond to an aggregation over a number of cells, including the cell of interest. In other cases the cell of interest will correspond to an aggregation over a number of perturbed cells.

In the former case the joint posterior distribution over the aggregated cells is required in order to calculate the posterior distribution over the cells' sum. But the complete independence implied by the uniform prior over T ensures that the distribution of the sum can be derived by a simple convolution of the marginal posteriors for the aggregated cells. Given the restricted range over the sum implied by a rounding scheme the convolution can be performed efficiently. In the case of a sample table we have a convolution of Hypergeometric density functions which is simply another Hypergeometric density function.

In the latter case a posterior distribution over the sum of aggregated cells can be calculated, but the intruder requires a posterior distribution over one of the individual cell counts. But after choosing, say, an improper uniform prior for each individual cell count, a posterior over the queried cell count can be derived. For rounded tables this can be performed efficiently. For sample tables it is not clear that exact inference can be performed efficiently.

But despite these possibilities for efficient, exact inference this paper is mainly concerned with inferences that are possible using naïve methods. Although it would be useful to be able to compare exact posterior probabilities with naïve probabilities, only naïve approaches will be investigated.

Naïve approaches

One possible naïve approach is to solve the lower and upper bounds for the queried cell, using methods similar to those of Dobra (2002), and to compute the naïve posterior probability distribution over the queried cell using the equation in Equation 5. In other words, the tables in R that have a single cell corresponding to the queried cell contribute directly to the calculation of the posterior, but only the feasible counts, derived from all tables in R, are summed over.

A more naïve approach, and the one used in this paper, is to perform the above, but with naïve bounds calculations. Taking into account all the interdependencies between cell counts can be computationally expensive for large base tables. A naïve approach avoids this complexity by examining the constraints placed on the queried cell count by each table individually. The bounds generated in this way are not guaranteed to be the tightest possible bounds, as can be achieved using the methods of Dobra and Fienberg (2000), and coupled with the naïve probability calculations might result in non-zero probabilities being assigned to non-feasible cell counts. But currently the more naïve approach is necessary for full tables with more than around 1000 cells. The following discussion considers these calculations for the specific attack scenario under consideration, a unique match across two samples. Here, the queried cells are those in the crosstabulation of the population over the variables common to both samples.

Bounds Calculations

The naïve bounds from the two samples are straightforward. A lower bound of 1 (for the relevant cell in the overlap margin) is implied by the unique match in the samples. An upper bound is given by N-n_{max}+1, where N is the known population size and n_{max} is the size of the larger sample.

For any marginal table that contains all the overlapping variables the constraints on the cell count can be found by generating the trivial lower and upper bounds and marginalizing to the overlapping variables. i.e. The value for the relevant cell in the overlap margin cannot be less than the sum of the minimum values for the set of cells whose sum equals the relevant cell value. Similarly the value for the relevant cell in the overlap margin cannot be greater than the sum of the maximum values for the set of cells whose sum equals the relevant cell value.

For any marginal table that contains a proper subset of the set of overlap variables we have a similar upper constraint on the relevant cell in the overlap margin. It cannot be greater than the corresponding cell in a smaller margin. But when it comes to the lower bound there is no obvious constraint (other than the obvious non-negativity constraint). The cell count might well be less than the corresponding cell in a smaller margin.

For example,

Let A + B = C for the exact counts, where A and B are cells in a table and C is the corresponding cell in a margin.

For perturbed counts where the marginal table contains all the overlapping variables (and we seek bounds on C given bounds on A and B),

$C_L >= A_L + B_L$

$C_U <= A_U + B_U$.

For perturbed counts where the table contains cells A and B, and the margin contains a proper subset of the overlapping variables (so we seek bounds on e.g. B given bounds on C),

$B_L <= C_L$

$B_U <= C_U$.

The third inequality, $B_L <= C_L$, never allows the obvious lower bound of 0 (or 1 in case of a unique match) to be tightened.
The trivial lower and upper bounds for the table counts stem directly from the rounding schemes used to generate them, or are equal to the table counts if the table has not been perturbed.
So, the trivial lower and upper bounds are generated for each table. The table of trivial upper bounds is marginalised to the set of variables that are common to the table and the overlap table. The upper bound (given the table) is computed as above. If the table contains all the overlap variables, then the table of lower bounds is marginalised and the lower bound computed as above. The lower bound (used for the calculation of the posterior) is the maximum of the lower bounds implied by the tables (and samples). The upper bound is the minimum of the upper bounds implied by the tables (and samples).

Example of analysis

An intruder has access to one or more sample tables, each relating to the same population. The samples have arbitrary, possibly overlapping, variable sets. The intruder can match records across the samples, and identify cases where the combined sample contains a unique match; that is, there is a set of records, one for each sample table, where all the overlapping sets of attributes match, and are themselves unique.

The intruder is interested in the probability that a unique match is a correct match. That is, the probability that a set of partial records all pertain to the same population unit. If a unique match in the samples has a corresponding count of 1 in the margin containing the overlapping variables, then the match would be correct. Counts greater than one would imply lower matching probabilities. The probability of a correct match can be derived

from a posterior distribution over the count in the relevant cell of the overlapping margin (subject to assumptions detailed later).

		Var3	
Var1	Var2	Level1	Level2
Level1	Level1	2	1
Level1	Level2	4	1
Level2	Level1	5	2
Level2	Level2	1	2

The above exact table might have the following two marginal samples released, with an overlap on Var2.

	Var2	
Var1	Level1	Level2
Level1	2	1
Level2	3	0

	Var3	
Var2	Level1	Level2
Level1	3	1
Level2	1	0

Pointwise (elementwise) multiplication of the sample tables gives all the possibilities for linkage between records in the samples,

		Var3	
Var1	Var2	Level1	Level2
Level1	Level1	6	2
Level1	Level2	1	0
Level2	Level1	9	3
Level2	Level2	0	0

and marginalizing to the overlapping variables (Var2) demonstrates that there is only one possibility for linking records in the samples containing Level2 of Var2.

Var2	
Level1	Level2
20	1

If the value in the corresponding margin in the exact table is 1, then the unique match in the samples must be a correct match. The existence of the matched records implies that the value in the corresponding margin of the exact table must be at least 1. For values greater than 1 the confidence in a correct match is much diminished. In this case we can see from the exact table that the value is actually 8. Thus there are 64 possible matches and only 8 that will be correct. Thus the 'true' marginal probability of a correct match is 1/8. NB. This assumes that all possible matchings are *a priori* equally likely. This is clearly a very big, and questionable, assumption, but seems to be made consistently within statistical disclosure control.

Given only the constraints implied by the samples and the known exact total (assumed to be known via known sampling fractions) the feasible range of values for the cell is 1 to 13. The posterior distribution of the cell value is calculated using the naïve method, and is shown below, with probabilities shown to 2 decimal places.

1	2	3	4	5	6	7	8	9	10	11	12	13
0.09	0.19	0.22	0.19	0.14	0.09	0.05	0.02	0.01	0.00	0.00	0.00	0.00

A randomly rounded (to base 3) version of the full table can be added to the release.

		Var3	
Var1	Var2	Level1	Level2
Level1	Level1	0	3
Level1	Level2	3	3
Level2	Level1	6	3
Level2	Level2	3	3

Recalculating the posteriors to take the new information into account gives the following posterior.

1	2	3	4	5	6	7	8	9	10	11	12	13
0	0	0	0.39	0.28	0.17	0.09	0.04	0.02	0.00	0.00	0	0

Although the rounded table is not taken into account for the calculation of posterior probabilities (as four rounded counts in the table correspond to the single cell in the overlap margin) it does tighten the bounds. The feasible range is now 4 to 11.

The fact that the minimum count is now 4 does not preclude a correct match in the samples. But we know that the marginal probability of a correct match for a count of c is simply $1/c$. The in intruder can generate the probability of a correct match by calculating the sum of the reciprocals of the feasible values, each weighted by the corresponding probability.

$0.39/4 + 0.28/5 + 0.17/6 + 0.09/7 + \ldots = 0.20$

This overestimates the true marginal probability of 0.125. (By 'true marginal probability' we mean the marginal probability of a match given full marginal tables rather than samples, assuming all possible matches are *a priori* equally likely.) This overestimation might be due to chance, or the naïve approach to probability calculations. Another possible cause is the initial search for (possibly degenerate) samples that had a unique match (that was undertaken for the purposes of this example). This example was chosen for illustrative purposes; no such search is required for the purposes of the following experiment.

An experiment

Various sets of tables were generated from the 1991 SAR. For each table a set of hypothetical release tables was generated (rounded and sampled tables on various sets of variables), as well as the two samples with overlapping variable sets used for the initial matching. The naïve method was applied to various hypothetical releases and used to rank the sample unique matches in order of their posterior probability of correctness. These were compared with the correct rankings derived from the known 'population'. Incremental releases were considered in order to identify types of table that were particularly informative.

Results

As many table releases were considered it is not possible to reproduce detailed results. But for the majority of table releases the rank correlations were highly significant. The only exceptions were cases with very few unique matches, in which cases the significance tests lacked power. Incremental additions of further tables to a table release tended to increase in the rank correlation. Tables that contributed directly to the calculation of the posteriors tended to affect the rank correlation more than those that only contributed to bounds calculations.

Conclusions

In general the ability to distinguish true from false linkages between records in microdata samples can be significantly enhanced by the co-presence of other tables, even when

using very naïve approaches. Under the naïve approach used, certain tables are significantly more informative than others. An intruder who uses a naïve Bayesian approach can avoid the computational costs associated with exact inference / Monte-Carlo methods and still have a useful tool for identifying linkages that are more likely than most to be correct. Similarly, very naïve Bayesian approaches might be useful for disclosure control. They can be used to identify the most 'obvious' risks without the usual computational burden.

References

Dobra, A. (2002). Statistical Tools for Disclosure Limitation in Multi-way Contingency Tables. PhD Thesis, Department of Statistics, Carnegie Mellon University.

Dobra, A. and Fienberg, S.E. (2000). Bounds for Cell Entries in Contingency Tables given Marginal Totals and Decomposable Graphs. Proceedings of the National Academy of Sciences, **97**, No.22, pp.11885-11892.

Smith, D. and Elliot, M. (2003). An Investigation of the Disclosure Risk Associated with the Proposed Neighbourhood Statistics. ONS Report, 2003.