MATHEMATICAL ENGINEERING
TECHNICAL REPORTS

Evaluation of per-record identification risk and swappability of records in a microdata set via decomposable models

Akimichi TAKEMURA and Yushi ENDO

METR 2006–17 March 2006

DEPARTMENT OF MATHEMATICAL INFORMATICS
GRADUATE SCHOOL OF INFORMATION SCIENCE AND TECHNOLOGY
THE UNIVERSITY OF TOKYO
BUNKYO-KU, TOKYO 113-8656, JAPAN

WWW page: http://www.i.u-tokyo.ac.jp/mi/mi-e.htm

Evaluation of per-record identification risk and swappability of records in a microdata set via decomposable models

Akimichi Takemura and Yushi Endo
Graduate School of Information Science and Technology
University of Tokyo

March, 2006

Abstract

We propose a strategy for disclosure risk evaluation and disclosure control of a microdata set based on fitting decomposable models of a multiway contingency table corresponding to the microdata set. By fitting decomposable models, we can evaluate per-record identification (or re-identification) risk of a microdata set. Furthermore we can easily determine swappability of risky records which does not disturb the set of marginals of the decomposable model. Use of decomposable models has been already considered in the existing literature. The contribution of this paper is to propose a systematic strategy to the problem of finding a model with a good fit, identifying risky records under the model, and then applying the swapping procedure to these records.

1 Introduction

In this paper we propose a systematic strategy of per-record identification risk and disclosure control of risky records of a microdata set by fitting decomposable models to a multiway contingency tables corresponding to the microdata. The first stage of our strategy consists of selecting decomposable models with a good fit to the data based on Akaike's information criterion (AIC). Since the number of decomposable models is large, we propose an algorithm to find locally optimum decomposable models. The second stage is to evaluate cell probabilities of sample unique records and to estimate the number of population uniques in the microdata set based on the chosen model. The third stage consists of disclosure control of risky records by swapping. We consider swapping which does not disturb the set of marginals corresponding to the chosen model.

In evaluating the disclosure risk of a given microdata set, the number of the population uniques among the sample unique records has been considered to an important

overall measure of the disclosure risk. Starting from Poisson-Gamma model ([1]) various models of random partitions have been proposed for estimating the number of population uniques. See a series of works of Hoshino ([12], [9], [10], [11]) and references therein. These models treat the sample unique records exchangeably and hence the estimated conditional probability of population uniqueness is common for every sample unique record. However some sample unique records are clearly more likely to be population uniques than other records, according to "rareness" of the records. If a sample unique has outlying observations or has very a rare combination of observed characteristics, it is likely to be a population unique. A simple descriptive method for evaluating per-record identification risk is to look at minimum unsafe combination of variables for a sample unique record ([17]).

More systematic way of evaluating the per-record identification risk is to model cell probabilities of the contingency table corresponding to a microdata set, where all the key variables of the microdata set are categorized and the joint frequencies of the key variables are counted. If the estimated cell probability of a sample unique cell is very small, then the sample unique is rare and risky. This approach was investigated in [14], [8], [3]. They used the standard log-linear models for cell probabilities of contingency tables.

In actual evaluation of disclosure risk, we often have to consider 10 or more possible key variables. Then the contingency table is large and sparse and the estimation of cell probabilities of standard log-linear models is not straightforward, except for decomposable models. In Section 2.2 we consider an example of a 8-way contingency table from 1990 U.S. Census of Population and Housing data. From the viewpoint of disclosure control this example is of moderate size but the contingency table corresponding to the microdata has more than 12 million cells.

Because of the computational difficulty Takemura [15] considered Lancaster-type additive modeling of cell probabilities. However in fitting additive models estimated cell probabilities often become negative, especially for empty cells. In this sense additive models are not satisfactory for estimating small cell probabilities, although they are useful for the purpose of relative evaluation of identification risks of sample unique cells.

Among the log-linear models, decomposable models are special in the sense that the maximum likelihood estimates of the cell probabilities can be explicitly written as ratios of products of marginal frequencies. Unlike other log-linear models, in a decomposable model cell probability of each cell can be separately estimated. This is a very attractive feature of decomposable model, because we are mainly interested in sample unique cells or other cells of small frequency. Furthermore model selection among decomposable models is relatively easy, because the maximized log likelihood and the degrees of freedom can be simply evaluated. For fitting other log-linear models, we need some iterative procedure such as iterative proportional scaling (see e.g. [6]). For large contingency tables iterative proportional scaling is computationally very intensive, because cell probability estimates of all the cells have to be stored in some form and updated in each iteration.

Estimation and diagnostics of a particular decomposable model is easy. However if the number m of key variables is large, there are many possible decomposable models. In Table 2 below, for our example of $m = 8$ key variables, there are more than 30 million

possible decomposable models. Finding the best fitting model among more than 30 million possible models is impractical. We propose to find several locally optimum models and choose one of these models.

Once a decomposable model with a good fit is obtained, we look at sample unique cells with very small estimated cell probabilities. If the cells are considered to be risky, it is desirable to perform some disclosure control measure to these cells. From the viewpoint of log-linear model, it is natural to consider swapping of these risky records in such a way that the swapping does not disturb the given set of marginals corresponding to the cliques of the decomposable model. This is based on the fact that the set of marginals constitutes the sufficient statistic of the model and swapping does not influence statistical inferences based on the model. Using the results of [18] we show that it is straightforward to determine whether a particular record is swappable and find another record for swapping if swapping is possible.

The organization of the paper is as follows. In Section 2 we summarize preliminary material and introduce our working example. In Section 3 we discuss fitting and selection of decomposable models. In Section 4 based on a chosen decomposable model we evaluate per-record identification risk. In Section 5 we perform swapping of risky records. Section 6 ends the paper with some concluding remarks.

2 Preliminaries and a working example

In this section we prepare notations on decomposable models and describe a working example analyzed in this paper.

2.1 Notations on decomposable models

We follow the notation of [13]. Let $\Delta = \{1, \ldots, m\}$ denote the set of the key variables. Each variable is denoted by $\delta \in \Delta$. We assume that all key variables are already discretized and let $\mathcal{I}_\delta = \{1, \ldots, I_\delta\}$ denote the set of categories of δ. Each cell is indexed by m indices $\boldsymbol{i} = (i_1, \ldots, i_m)$ and the set of the cells is the direct product $\mathcal{I} = \prod_{\delta \in \Delta} \mathcal{I}_\delta$. The frequency of cell $\boldsymbol{i}$ is denoted by $n(\boldsymbol{i})$.

Let $a \subset \Delta$ be a subset of variables. Then an a-marginal cell $\boldsymbol{i}_a$ of $\boldsymbol{i} = (i_1, \ldots, i_m)$ is defined as $\boldsymbol{i}_a = (i_\delta)_{\delta \in a}$. The set of a-marginal cells is $\mathcal{I}_a = \prod_{\delta \in a} \mathcal{I}_\delta$. The marginal frequency of a-marginal cell $\boldsymbol{i}_a$ is written as

$$n(\boldsymbol{i}_a) = \sum_{\boldsymbol{j} : \boldsymbol{j}_a = \boldsymbol{i}_a} n(\boldsymbol{j}),$$

where $\boldsymbol{j}_a = \boldsymbol{i}_a$ means $i_k = j_k,\ \forall k \in a$. Let $n = \sum_{\boldsymbol{i} \in \mathcal{I}} n(\boldsymbol{i})$ denote the sample size (number of records) of the microdata set. We denote the relative frequency of a cell $\boldsymbol{i}$ and a marginal cell $\boldsymbol{i}_a$ by

$$r(\boldsymbol{i}) = \frac{n(\boldsymbol{i})}{n}, \quad r(\boldsymbol{i}_a) = \frac{n(\boldsymbol{i}_a)}{n}.$$

We use the same notation for cell probabilities $p(\boldsymbol{i})$, $p(\boldsymbol{i}_a)$, etc.

Consider a graph $G = (\Delta, E)$ with the set of vertices Δ and the set of edges E. Let $\mathcal{C}$ denote the set of (maximal) cliques. For a subset $a \subset \Delta$ let $\mu_a : \mathcal{I} \to R$ denote a function of $\boldsymbol{i}$ which only depends on the marginal cell $\boldsymbol{i}_a$, i.e. $\mu_a(\boldsymbol{i}) = \mu_a(\boldsymbol{i}_a)$. Then the graphical model associated with G specifies the cell probability $p(\boldsymbol{i})$ as

$$\log p(\boldsymbol{i}) = \sum_{a \in \mathcal{C}} \mu_a(\boldsymbol{i}_a). \tag{1}$$

A graph G is *chordal* (*decomposable*, *triangulated*), if every cycle of length $l \geq 4$ has a chord. A graphical model with a chordal G is called a *decomposable* model. For a decomposable model, the cliques can be ordered to satisfy the running intersection property:

$$\begin{aligned} &\text{(RIP)} \quad \text{For each } 2 \leq j \leq m, \text{ there exists } 1 \leq k \leq j-1, \text{ such that} \\ &\qquad S_j = C_j \cap (C_1 \cup C_2 \cup \cdots \cup C_{j-1}) \subset C_k. \end{aligned}$$

An ordering $(C_1, \ldots, C_m)$ satisfying RIP is called a *perfect sequence*. $S_2, \ldots, S_m$ are minimal vertex separators of G. The number of times a minimal vertex separator S appears in any perfect sequence is the same and called the *multiplicity* of S. We denote the multiplicity of S by $\nu(S)$. $\mathcal{S}$ denotes the set of minimal vertex separators. In the following we simply say "separator" to mean a minimal vertex separator.

The maximum likelihood estimate (MLE) of a decomposable model is explicitly written as

$$\hat{p}_{\mathrm{MLE}}(i) = \begin{cases} \dfrac{\prod_{C \in \mathcal{C}} r(i_C)}{\prod_{S \in \mathcal{S}} r(i_S)^{\nu(S)}}, & \text{if } \ r(i_C) > 0, \ \forall C \in \mathcal{C}, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

The degrees of freedom is also simply written (Proposition 4.35 of [13]).

$$\sum_{C \in \mathcal{C}} \prod_{\delta \in C} I_\delta - \sum_{S \in \mathcal{S}} \nu(S) \prod_{\delta \in S} I_\delta. \tag{3}$$

Hence AIC for model selection is also easily computed.

$$\mathrm{AIC} = -2 \times (\text{log likelihood}) + 2 \times (\text{degrees of freedom}). \tag{4}$$

In Table 1 we list the number of graphical models and the number of decomposable models for m-way contingency tables up to $m = 8$. We see that the number of decomposable models increases very fast with m. The number in the parentheses for the decomposable model indicates the number of chordal graphs of m vertices after identification of isomorphic graphs, i.e., we do not distinguish graphs which can be obtained by relabeling of vertices. Based on [4] we provide a list of non-isomorphic chordal graphs for $m \leq 8$ in [5]. Given a list of non-isomorphic chordal graphs we can pick a decomposable model by choosing an graph from the list and arbitrary assigning a variable to each vertex of the graph.

Table 1: Number of graphical models and decomposable models

m	graphical	decomposable
2	2	2 (2)
3	8	8 (4)
4	64	61 (10)
5	1024	820 (27)
6	32768	18154 (96)
7	2097152	617675 (469)
8	268435456	30888596 (3734)

2.2 A working example

In this paper we apply our strategy to a test data set from 1990 U.S. Census of Population and Housing Public Use Microdata Samples. We subsampled $n = 9809$ individuals from the state of Washington and chose $m = 8$ variables for our experiment.

1. Relationship (14 categories)	2. Sex (2 categories)
3. Age (91 categories)	4. Marital status (5 categories)
5. Place of birth (14 categories)	6. Spouse present/absent (7 categories)
7. Own child (2 categories)	8. Age of own child (5 categories)

The population size of the state of Washington is about $N = 4,867,000$. The dataset can be viewed as a 8-way contingency table of the type

$$14 \times 2 \times 91 \times 5 \times 14 \times 7 \times 2 \times 5$$

with approximately 12.5 million cells (more exactly 12,485,200 cells). We see that the contingency table is very sparse with only $n = 9809$ counts in 12.5 million cells. We took these $m = 8$ variables from a PUMS data set without further global recoding. For example we used the age itself with 91 categories. This is somewhat unrealistic for evaluation of disclosure risk. On the other hand there are other possible key variables in the original PUMS data set.

It should be noted that although the (formal) total number of cells 12,485,200 is very large, the effective total number should be much smaller because of structural zeros. For example there is no age of own child if there is no own child. In this case the age of own child is coded as N/A in the original data set. Also there is an obvious relation between age and marital status. In this paper we ignore the effect of structural zeros. See Section 6 for more discussion.

For reference we show first few lines of 9809×8 data matrix.

```
00,0,17,4,10,6,0,0
00,0,17,4,52,6,0,0
```

```
00,0,18,0,23,1,0,0
00,0,18,0,24,1,0,0
00,0,18,0,51,1,0,0
```

The frequencies of the cell sizes (size indices, frequency of frequencies) of this data set is given as follows. The table shows that there are 2243 cells of frequency 1, 524 cells of frequency 2, etc.

Cell size	1	2	3	4	5	6	7	8	9	10	11 $\leq$
Frequency	2243	524	275	132	104	60	59	34	46	19	124

We are interested in estimating the number of population uniques among 2243 sample uniques and evaluate which sample record is particularly risky. As a preliminary analysis, we fitted Ewens model, Pitman model and Lancaster-type additive model. The estimates of the number of population uniques of these models are as follows.

Ewens model: 5.9, Pitman model: 214.0, additive model: 252.1.

3 Selection of decomposable models

The first step of our strategy is to choose a decomposable model which fits the data. As shown in Table 1 the number of possible decomposable models grow very fast as the number of variables m increases. For $m \leq 8$ we can use the list of non-isomorphic chordal graphs available at [5]. We present the following Algorithm 1 to obtain locally best decomposable model in terms of AIC. Application of Algorithm 1 to the data set of our working example is summarized in Table 2 below.

In our algorithm we add or subtract an edge to (or from) a chordal graph to move to another chordal graph and evaluate AIC. It outputs a model with locally minimum AIC. We can apply our algorithm from various initial models and compare these locally best models to obtain approximately a globally best model.

Notations of Algorithm 1 is as follows. $G = G(V, E) = G(V, E_G)$ is a graph with the set of vertices V and the set of vertices E. $M(G)$ denotes the graphical model associated with G. $E(C_m)$ denotes the set of edges of the complete graph with m vertices.

In Step 1 we choose an initial model randomly from the list of non-isomorphic decomposable models([4], [5]). Then we randomly label the vertices to obtain a decomposable model. We will discuss random generation of initial models for $m > 8$ in Algorithm 2 below.

In Step 2 we choose the candidate for next decomposable model. We add or subtract an edge and determine whether the resulting graph is chordal. If it is chordal we evaluate its AIC. For evaluating AIC we need to obtain the set of cliques and the set of separators. Chordality of a graph is determined by obtaining a perfect elimination scheme and the set

of cliques and the separators are obtained by "Maximum cardinality search" algorithm ([2]).

Algorithm 1 Model selection of decomposable models.
Input: Microdata D, List of non-isomorphic chordal graphs $\mathcal{L}_m$ with m vertices
Output: Model M with local minimum AIC.

1. Choose a chordal graph from $H \in \mathcal{L}_m$ at random;
   Label vertices of H at random and obtain a chordal graph G_{next};
   $A_{\text{next}} \leftarrow$ AIC of $M(G_{\text{next}})$;
2. **while** $f = true$ **do**
   $f = false$;
   $G \leftarrow G_{\text{next}}$;
   for each $e \in E(C_m)$ **do**
   if $e \in E_G$ **then**
   $G' \leftarrow G(V, G(E) \setminus e)$
   else
   $G' \leftarrow G(V, G(E) \cup e)$;
   if G' is chordal **then**
   $A' \leftarrow$ AIC of $M(G')$;
   if $A' < A_{\text{next}}$ **then**
   $G_{\text{next}} \leftarrow G'$;
   $A_{\text{next}} \leftarrow A'$;
   $f \leftarrow true$;
3. Output $M(G)$;

For $m > 8$ we can propose the following algorithm to generate an initial decomposable model to replace Step 1 of Algorithm 1. Given a chordal graph G with m vertices, we can obtain a chordal graph G' with $m + 1$ vertices by adding the $m + 1$'st vertex and connecting it to a subset of one clique C of G. Since a chordal graph possesses a perfect sequence of cliques, the above recursive procedure generates all chordal graphs. The following Algorithm 2 outputs the set of cliques of a random chordal graph. Note that the probability distribution on random choices in the algorithm is not specified and the distribution of the output is not necessarily the uniform distribution over the set of chordal graphs with m vertices.

Algorithm 2 "Random" chordal graph with m vertices.
Input: m
Output: Set of cliques of a random chordal graph with m vertices.

1. Initialize $\mathcal{C} = \emptyset$;
2. **for** $j \leftarrow 1$ **until** m **do**
   Flip a coin;
   if heads **then**
   $\mathcal{C} \leftarrow \mathcal{C} \cup \{\{j\}\}$
   else choose a member $C \in \mathcal{C}$ and a subset $C' \subset C$ at random;

if $C = C'$ **then**
$C \leftarrow C \cup \{j\}$
else
$\mathcal{C} \leftarrow \mathcal{C} \cup \{C' \cup \{j\}\}$;
3. Output $\mathcal{C}$;

4 Per-record identification risk and estimate of the number of population uniques

When a good fitting decomposable model is chosen we can estimate the cell probability of a sample unique cell by MLE (2). Then a natural estimate of the conditional probability that the sample unique cell $\boldsymbol{i}$ is also a population unique is given as

$$(1 - \hat{p}_{\mathrm{MLE}}(\boldsymbol{i}))^{N-n}, \tag{5}$$

where N is the population size and n is the sample size. (5) is the estimated probability that none of the remaining $N - n$ individuals in the population fall into cell $\boldsymbol{i}$, under the assumption that individuals fall into cells independently from each other according to the estimated probability distribution. The number of population uniques in the sample can be estimated as

$$\sum_{\boldsymbol{i}:\text{sample unique}} (1 - \hat{p}_{\mathrm{MLE}}(\boldsymbol{i}))^{N-n}.$$

In Table 2 we show two models with smallest values of AIC by applying Algorithm 1 100 times to our example. Algorithm 1 converged after a few transitions and it seems to be very practical. These two models were also most frequently obtained from Algorithm 1. In both models, the separator $\{6\}$ has multiplicity 2 as indicated by the repetition in the table. The estimated numbers of population uniques (48.867, 40.51) are between those of Ewens model and Pitman model and seem to be reasonable. The variable 6 (Spouse present/absent) is contained in many cliques, which can be explained by its high correlation with other variables and yet small degrees of freedom. On the other hand variable 5 (Place of birth) is contained in a single clique (i.e. it is a simplicial vertex), which is also reasonable.

Furthermore the sample uniques with very small estimated cell probabilities ($p(i) \leq 10^{-8}$) are common to these two models. We might consider some disclosure control measure for about 20 sample uniques with estimated cell probability less than 10^{-7}.

5 Swappability of risky records

In Table 2 two records have the estimated cell probability of less than 10^{-8}. They probably need some disclosure control. In this paper we propose to swap some observations of these records with other records of the data set. Since we have found a decomposable model

Table 2: Chosen models

	Model 1	Model 2
Number of times chosen	11	7
AIC/2	13869.07	13984.97
log likelihood	−12141.07	−12013.97
degrees of freedom	1728	1971
estimated # of population uniques	48.867	40.515
cliques	{1,2,6},{1,6,7},{2,6,8}, {3,6,7},{4,6},{5,6}	{1,6,7},{3,6,7},{1,6,8}, {2,8},{4,6},{5,6}
separator	{1,6},{2,6},{6,7},{6},{6}	{1,6},{6,7},{6},{6},{8}
cell probability estimates	frequencies	frequencies
10^{-2} to 10^{-3}	0	0
10^{-3} to 10^{-4}	352	351
10^{-4} to 10^{-5}	1092	1117
10^{-5} to 10^{-6}	599	600
10^{-6} to 10^{-7}	179	158
10^{-7} to 10^{-8}	19	15
10^{-8} to 10^{-9}	2	2
10^{-9} to 10^{-10}	0	0

with a good fit, it is desirable to swap the observations such that the marginal frequencies for the cliques of the chosen model is not disturbed. In [18] we give some necessary and sufficient conditions for swappability of a particular sample unique record with some other record without disturbing a given set of marginals.

For a decomposable model, a simple method for searching another record for swapping can be described as follows. Let $\boldsymbol{i}$ be a sample unique record, such that we want to swap some observations of this record with another record. Let $\mathcal{C}$ be the set of cliques of a chosen model and let $\mathcal{S}$ denote the set of minimal vertex separators. Write each separator S as the intersection of two cliques $S = C \cap C'$. We consider all triples (C, C', S) such that $S = C \cap C'$. For example in Model 1 in Table 2 all possible ways of writing separators are as follows.

$$
\begin{aligned}
\{1,6\} &= \{1,2,6\} \cap \{1,6,7\},\\
\{2,6\} &= \{1,2,6\} \cap \{2,6,8\},\\
\{6,7\} &= \{1,6,7\} \cap \{3,6,7\},\\
\{6\} &= \{1,2,6\} \cap \{3,6,7\} = \{1,2,6\} \cap \{4,6\} = \{1,2,6\} \cap \{5,6\}\\
&= \{1,6,7\} \cap \{2,6,8\} = \{1,6,7\} \cap \{4,6\} = \{1,6,7\} \cap \{5,6\}\\
&= \{2,6,8\} \cap \{3,6,7\} = \{2,6,8\} \cap \{4,6\} = \{2,6,8\} \cap \{5,6\}
\end{aligned}
$$

$$= \{3,6,7\} \cap \{4,6\} = \{3,6,7\} \cap \{5,6\} = \{4,6\} \cap \{5,6\}.$$

For a particular sample unique record $\boldsymbol{i}$, we search other records $\boldsymbol{j} \neq \boldsymbol{i}$ such that for some (C, C', S) we have

$$\boldsymbol{i}_S = \boldsymbol{j}_S, \quad \boldsymbol{i}_C \neq \boldsymbol{j}_C, \quad \boldsymbol{i}_{C'} \neq \boldsymbol{j}_{C'} \tag{6}$$

If we find some $\boldsymbol{j}$ and some (C, C', S) such that (6) holds, then we can swap some observations between $\boldsymbol{i}$ and $\boldsymbol{j}$.

We applied this procedure to 50 sample unique records with small estimated cell probabilities in Table 2. For both models of Table 2 this procedure quickly found other records for swapping for most of 50 records, including the two records with the estimated cell probability of less than 10^{-8}. Therefore this procedure seems to work very well in practice.

Note that (6) is a sufficient condition for swappability between $\boldsymbol{i}$ and $\boldsymbol{j}$ for a decomposable model. For a full statement of necessary and sufficient conditions for general hierarchical model see Section 3 of [18].

6 Concluding remarks

In this paper we proposed a systematic strategy for disclosure risk evaluation and disclosure control of microdata set by fitting decomposable models. We have restricted our attention to decomposable models in view of computational convenience. Clearly it is desirable to consider other hierarchical models such as the model containing all two-factor interaction terms. Simpler hierarchical model might give a better fit than more complicated decomposable model. One strategy we can try is to look for hierarchical models which improves the fit around a locally best decomposable model.

We have used AIC for evaluating the fit of the model. Theoretically AIC is justified for large sample size. In disclosure control problems we are dealing with large and sparse tables and from theoretical viewpoint use of AIC is not justified . However in practice it is simple and seems to work reasonably well. It is of interest to investigate other methods of model selection for evaluating the fit of various models.

In microdata sets of official statistics, there are large number of structural zeros due to various logical relations between key variables. In principle we should list all the logical relations and specify structural zeros before fitting a model. But this is very cumbersome. Also the calculation of degrees of freedom of a model becomes complicated. It is desirable to develop some practical methods to deal with structural zeros in some automatic way.

If we want to swap some observations from a sample unique record $\boldsymbol{i}$ and if we can find many other records $\boldsymbol{j}$ for swapping, it might be desirable to use $\boldsymbol{j}$ which is close to $\boldsymbol{i}$ in some sense. In [16] we considered swapping of observations between close records by introducing distance function between records.

Acknowledgment The approach of this paper was suggested in a talk by Stephen Fienberg [7] at University of Tokyo in May 2003 and we are very grateful to his insights.

It took us a long time to implement the whole strategy based on his suggestions.

References

[1] Jelke G. Bethlehem, Wouter J. Keller, and Jeroen Pannekoek. Disclosure control of a microdata. *Journal of the American Statistical Association*, 85:38–45, 1990.

[2] Jean R. S. Blair and Barry Peyton. An introduction to chordal graphs and clique trees. In *Graph theory and sparse matrix computation*, volume 56 of *IMA Vol. Math. Appl.*, pages 1–29. Springer, New York, 1993.

[3] Elsayed A.H. Elamir. Analysis of re-identification risk based on log-linear models. In *Lecture Notes in Computer Science, Volume 3050*, pages 273 – 281. Springer, 2004.

[4] Yushi Endo. Algorithms for enumeration of decomposable models. Bachelor's thesis, Department of Mathematical Engineering and Information Physics, University of Tokyo, 2004.

[5] Yushi Endo and Akimichi Takemura. List of chordal graphs up to 8 vertices. `http://www.stat.t.u-tokyo.ac.jp/~takemura/decomposable.html`, 2004.

[6] Yushi Endo and Akimichi Takemura. Iterative proportional scaling via decomposable submodels for contingency tables. Technical Report METR 2006-16, University of Tokyo, 2006. Submitted for publication.

[7] Stephen E. Fienberg. Log-linear models and computational algebra: old wine in new bottles?, May 27 2003. Talk at the joint statistics seminar at University of Tokyo.

[8] Stephen E. Fienberg and Udi E. Makov. Confidentiality, uniqueness, and disclosure limitation for categorical data. *Journal of Official Statistics*, 14:385–397, 1998.

[9] Nobuaki Hoshino. Applying Pitman's sampling formula to microdata disclosure risk assessment. *Journal of Official Statistics*, 17(4):499–520, 2001.

[10] Nobuaki Hoshino. Random clustering based on the conditional inverse Gaussian-Poisson distribution. *J. Japan Statist. Soc.*, 33(1):105–117, 2003.

[11] Nobuaki Hoshino. Engen's extended negative binomial model revisited. *Ann. Inst. Statist. Math.*, 57(2):369–387, 2005.

[12] Nobuaki Hoshino and Akimichi Takemura. Relationship between logarithmic series model and other superpopulation model useful for microdata disclosure risk assessment. *Journal of the Japan Statistical Society*, 28:125–134, 1998.

[13] Steffen L. Lauritzen. *Graphical Models*. Oxford University Press, Oxford, 1996.

[14] C. J. Skinner and D. J. Holmes. Estimating the re-identification risk per record in microdata. *Journal of Official Statistics*, 14:361–372, 1998.

[15] Akimichi Takemura. Evaluation of per-record identification risk by additive modeling of interaction for contingency table cell probabilities. In *IASS Proceedings - SEOUL 2001, International Association of Survey Statisticians, The International Statistical Institute*, pages 220–235, 2002.

[16] Akimichi Takemura. Local recording and record swapping by maximum weight matching for disclosure control of microdata sets. *Journal of Official Statistics*, 18(2):275–289, 2002.

[17] Akimichi Takemura. Minimum unsafe and maximum safe sets of variables for disclosure risk assessment of individual records in a microdata set. *Journal of the Japan Statistical Society*, 32:107–117, 2002.

[18] Akimichi Takemura and Hisayuki Hara. Conditions for swappability of records in a microdata set when some marginals are fixed. Technical Report METR 2006-18, University of Tokyo, 2006. Submitted for publication.