MATHEMATICAL ENGINEERING
TECHNICAL REPORTS

Conditions for swappability of records in a microdata set when some marginals are fixed

Akimichi TAKEMURA and Hisayuki HARA

METR 2006–18 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

Conditions for swappability of records in a microdata set when some marginals are fixed

Akimichi Takemura
Graduate School of Information Science and Technology
University of Tokyo
and
Hisayuki Hara
Department of Geosystem Engineering
University of Tokyo

March, 2006

Abstract

We consider swapping of two records in a microdata set for the purpose of disclosure control. We give some necessary and sufficient conditions that some observations can be swapped between two records under the restriction that a given set of marginals are fixed. We also give an algorithm to find another record for swapping if one wants to swap out some observations from a particular record. Our result has a close connection to the construction of Markov bases for contingency tables with given marginals.

Keywords and phrases: decomposable model, disclosure control, graphical model, hierarchical model, Markov basis, primitive move.

1 Introduction

In statistical disclosure control of microdata sets, swapping of observations among records is considered to be a convenient disclosure control technique, especially because it preserves one-dimensional marginals. Data swapping was introduced by Dalenius and Reiss [1982] and Schlörer [1981]. Takemura [2002] considered optimal pairing of close records of a microdata set to perform swapping. As explained in Dobra [2003] and Dobra and Sullivant [2004], swapping has a close connection to the theory of Markov bases for contingency tables. See Willenborg and de Waal [2001] for a review of disclosure control techniques for microdata sets.

Suppose that a statistical agency is considering to grant access to a microdata set to some researchers and the data set contains some rare and risky records. We consider the case that all variables of the data set have been already categorized. Swapping of observations is one of the useful techniques of protecting these records. If some marginals from the data set have been already published, it is desirable to perform the swapping in such a way that the swapping does not disturb the published marginal frequencies. Therefore it is important to determine, whether it is possible to perform swapping of risky records under the restriction that some marginal are fixed. See Takemura and Endo [2006] for a realistic example of the need for swapping.

Feasibility of swapping under the restriction that some marginal are fixed depends on the set of fixed marginals. We here illustrate this point by a simple hypothetical example. Suppose that a microdata set contains the following two records.

sex	age	occupation	residence
male	55	nurse	Tokyo
female	50	police officer	Osaka

If we swap "occupation" among these two records we obtain

sex	age	occupation	residence
male	55	police officer	Tokyo
female	50	nurse	Osaka

By this swapping the one-dimensional marginals are preserved, but the two-dimensional marginal of {age, occupation} is disturbed. If we swap both age and occupation we obtain

sex	age	occupation	residence
male	50	police officer	Tokyo
female	55	nurse	Osaka

and {age, occupation}-marginal is also preserved.

This simple example shows that observations can be freely swapped if we fix only the one-dimensional marginals, but some observations have to be swapped together to keep two-dimensional marginals fixed.

In fact if all two-dimensional marginals are fixed, then it is impossible to swap observations between any two records without disturbing at least one of the two-dimensional marginals. This is because if some observations are swapped and some observations are not swapped between two records, then the two-dimensional marginal of a swapped variable and a non-swapped variable is disturbed. This fact is clarified in a general form in Theorem 3.1 in Section 3.1.

Actually there is a possibility of swapping observations involving more than two records to keep all two-dimensional marginals fixed. We present an example of this possibility in Section 4. Swapping among more than two records is closely related to higher degree moves of Markov bases for contingency tables. It is well known that Markov basis involving higher degree moves is very complicated (e.g. Aoki and Takemura [2003]).

In this paper we consider swapping between two records only and we give some necessary and sufficient conditions for swappability of two records such that a given set of marginals are fixed. We also give a practical algorithm to find another record for swapping if one wants to swap out some observations from a particular record. Our conditions are conveniently described in terms decompositions by minimal vertex separators of a graphical model generated by the set of marginals. Results of the present paper are successfully applied in Takemura and Endo [2006] to check swappability of risky records in a microdata set of a substantial size.

The organization of this paper is as follows. In Section 2 we summarize notations and present some preliminary results including the equivalence of swapping between two records and a primitive move of a Markov basis. In Section 3 we give some necessary and sufficient conditions for swappability of two records. We also give an algorithm to find another record for swapping for a particular record. Some discussions are given in Section 4. Technical details are postponed to Appendix.

2 Preliminaries

In this section we first setup appropriate notations and summarize some preliminary results for this paper. Consider an $n \times k$ microdata set X consisting of observations on k variables for n individuals (records). As mentioned above we assume that the variables have been already categorized. Therefore we can identify the microdata set with a k-way contingency table, if we ignore the labels of the individuals. Concerning contingency tables, we mostly follow the notation in Dobra [2003] and Dobra and Sullivant [2004]. $\mathbf{n}$ denotes a k-way contingency table. For positive integer m, $\{1, \ldots, m\}$ is denoted by $[m]$. Let $\Delta = [k] = \{1, \ldots, k\}$ denote the set of variables. The cells of the contingency table are denoted by $i = (i_1, \ldots, i_k) \in \mathcal{I} = [I_1] \times \cdots \times [I_k]$. Each record of the microdata set falls into some cell i. $n(i)$ denotes the frequency of cell i. If $n(i) = 1$, we say that the record falling into cell i is a *sample unique record*.

For a subset $D \subset \Delta$ of variables, the D-marginal $\mathbf{n}_D$ of $\mathbf{n}$ is the contingency table with marginal cells $i_D \in \mathcal{I}_D = \prod_{j \in D}[I_{i_j}]$ and entries given by

$$n_D(i_D) = \sum_{i_{D^C} \in \mathcal{I}_{D^C}} n(i_D, i_{D^C}).$$

Here we are denoting $i = (i_D, i_{D^C})$ by ignoring the order of the indices.

Let E be a non-empty proper subset of Δ. For two records of X falling into cells $i = (i_E, i_{E^C})$ and $j = (j_E, j_{E^C})$, $i \neq j$, swapping of i and j with respect to $E \subset \Delta$, or more simply E-swapping, means that these records are changed as

$$\{(i_E, i_{E^C}), (j_E, j_{E^C})\} \to \{(i_E, j_{E^C}), (j_E, i_{E^C})\}. \tag{1}$$

Note that E-swapping is equivalent to E^C-swapping. Also note that if $i_E = j_E$ or $i_{E^C} = j_{E^C}$, then swapping in (1) results in the same set of records. Therefore (1) results in a

different set of records if and only if

$$i_E \neq j_E \quad \text{and} \quad i_{E^C} \neq j_{E^C}. \tag{2}$$

From now on we say that E-swapping is *effective* if it results in a different set of records.

We now ask when E-swapping fixes D-marginals. D-marginals are fixed by E-swapping if and only if one of the following four conditions holds.

$$\text{i) } D \subset E, \quad \text{ii) } D \subset E^C, \quad \text{iii) } i_{E\cap D} = j_{E\cap D}, \quad \text{iv) } i_{E^C\cap D} = j_{E^C\cap D}. \tag{3}$$

It is obvious that if one of the conditions holds, then D-marginals are not altered. On the other hand assume that all four conditions do not hold. Let $D_1 = D\cap E$ and $D_2 = D\cap E^C$. These are non-empty because i) and ii) do not hold. Furthermore $i_{D_1} \neq j_{D_1}$ and $i_{D_2} \neq j_{D_2}$ because iii) and iv) do not hold. Let $i_D = (i_{D_1}, i_{D_2})$. Then $n_D(i_D) = n_D(i_{D_1}, i_{D_2})$ is decreased by 1 by this swapping and this particular D-marginal changes.

So far we have only considered one marginal D. We need to consider a set of marginals $\mathcal{D} = \{D_1, \ldots, D_r\}$. For simplicity throughout this paper we assume $\Delta = \cup_{s=1}^r D_s$. If $\cup_{s=1}^r D_s$ is a proper subset of Δ, we can simply replace Δ by $\cup_{s=1}^r D_s$, because there is no restriction on frequency distributions involving variables in $(\cup_{s=1}^r D_s)^C$. We investigate conditions for swapping two records such that all marginals in $\mathcal{D}$ are fixed. Note that a smaller marginal can be computed by further summation of frequencies of a larger marginal. This implies that in $\mathcal{D}$ we only need to consider $D_1, \ldots, D_r$, such that there is no inclusion relation between them, i.e. $\mathcal{D}$ is an "antichain" (Klain and Rota [1997]). Any antichain $\mathcal{D}$ is a generating class of a hierarchical model for the contingency table (Lauritzen [1996]).

A hierarchical model with a generating class $\mathcal{D}$ is graphical if $\mathcal{D}$ coincides with a set of (maximal) cliques of a graph G with vertex set Δ. A graphical model is decomposable if G is a chordal graph.

Given a generating class $\mathcal{D}$, we define a graph $G^{\mathcal{D}}$ *generated* by $\mathcal{D}$ as follows. The vertex set of $G^{\mathcal{D}}$ is Δ. We put an edge between $s, t \in \Delta$ if and only if there exists $D \in \mathcal{D}$ such that $\{s, t\} \subset D$. Note that the graphical model associated with $G^{\mathcal{D}}$ is the smallest graphical model containing the hierarchical model with the generating class $\mathcal{D}$.

An integer array $\mathbf{f} = \{f(i)\}_{i\in\mathcal{I}}$ is a move for $\mathcal{D}$ if $f_D(i_D) \equiv 0$ for all $D \in \mathcal{D}$. $\mathbf{f}$ is a *primitive move* for $\mathcal{D}$ if it is a move for $\mathcal{D}$ and furthermore if two entries of $\mathbf{f}$ are 1, two entries are -1 and the other entries are 0. Adding a move $\mathbf{f}$ to $\mathbf{n}$, or applying $\mathbf{f}$ to $\mathbf{n}$, obviously does not alter the D-marginal for every $D \in \mathcal{D}$. It is intuitively clear that a primitive move and swapping of observations of two records are equivalent. In fact Dobra [2003] does not distinguish these two. However there is at least a conceptual difference between them, because a move is defined for a given set of marginals $\mathcal{D}$ whereas E-swapping is defined only in terms of two records and a subset E. We give a proof of this equivalence in Appendix.

3 Necessary and sufficient conditions of swappability

In this section we give some necessary and sufficient conditions for swappability of observations between two records. In particular in Theorem 3.1 we state a necessary and sufficient condition in terms of an induced subgraph of $G^{\mathcal{D}}$, which is convenient for application. Then we describe a practical algorithm to find another record for swapping for a particular record.

3.1 Swappability between two records

In (3) we have already given a necessary and sufficient condition for E-swapping to fix D-marginals. However (3) is not very useful for considering simultaneous fixing of marginals in $\mathcal{D} = \{D_1, \ldots, D_r\}$.

For clear argument it is better to distinguish variables which are common in two records and variables which have different values in two records. Note that if some variable has the same value in two records, swapping or no swapping of the variable do not make any difference. Therefore we should only look at variables taking different values in two records. Let

$$\bar{\Delta} = \{s \mid i_s \neq j_s\} \tag{4}$$

denote the set of variables taking different values in two records. Note that (2) holds if and only if

$$E \cap \bar{\Delta} \neq \emptyset \quad \text{and} \quad E^C \cap \bar{\Delta} \neq \emptyset. \tag{5}$$

Therefore E-swapping effective if and only if $E \cap \bar{\Delta} \neq \emptyset$ and $E^C \cap \bar{\Delta} \neq \emptyset$. In particular $\bar{\Delta}$ has to contain at least two elements, because if $\bar{\Delta}$ has less than two elements swapping between i and j can not result in a different set of records.

We now show the following lemma. The following lemma says that the variables in $\bar{\Delta} \cap D$ have to be swapped simultaneously or otherwise stay together in order not to disturb D-marginals.

Lemma 3.1. *An effective E-swapping fixes D-marginals if and only if $\bar{\Delta} \cap D \subset E$ or $\bar{\Delta} \cap D \subset E^C$ under (5).*

Proof. We have to check that one of the four conditions in (3) holds if and only if $\bar{\Delta} \cap D \subset E$ or $\bar{\Delta} \cap D \subset E^C$.

Assume that one of the four conditions in (3) holds. If $D \subset E$, then $\bar{\Delta} \cap D \subset E$. Similarly if $D \subset E^C$, then $\bar{\Delta} \cap D \subset E^C$. Now suppose $i_{E \cap D} = j_{E \cap D}$. Then

$$\emptyset = \bar{\Delta} \cap (E \cap D) = (\bar{\Delta} \cap D) \cap E \qquad \Rightarrow \qquad \bar{\Delta} \cap D \subset E^C.$$

Similarly if $i_{E^C \cap D} = j_{E^C \cap D}$ then $\bar{\Delta} \cap D \subset E$.

Conversely assume that $\bar{\Delta} \cap D \subset E$ or $\bar{\Delta} \cap D \subset E^C$. In the former case $\bar{\Delta} \cap D \cap E^C = \emptyset$ and this implies iv) $i_{E^C \cap D} = j_{E^C \cap D}$. Similarly in the latter case iii) $i_{E \cap D} = j_{E \cap D}$ holds. □

In the above lemma, E is given. Now suppose that two records i, j and a marginal D is given and we are asked to find a non-empty proper subset $E \subset \Delta$ such that E-swapping is effective and fixes D-marginals. As a simple consequence of Lemma 3.1 we have the following lemma.

Lemma 3.2. *Given two records i, j and $D \subset \Delta$, we can find $E \subset \Delta$ such that E-swapping is effective and fixes D-marginals if and only if $\bar{\Delta} \cap D^C \neq \emptyset$ and $|\bar{\Delta}| \geq 2$.*

Proof. If $\bar{\Delta} \cap D^C \neq \emptyset$ and $|\bar{\Delta}| \geq 2$, then choose $s \in \bar{\Delta} \cap D^C$ and let $E = \{s\}$ to be a one-element set. Then E satisfies the requirement.

If $|\bar{\Delta}| \leq 1$, there is no E-swapping resulting in a different set of records as mentioned above. On the other hand if $\bar{\Delta} \cap D^C = \emptyset$ or $\bar{\Delta} \subset D$, then by Lemma 3.1 $\bar{\Delta} \subset E$. But this contradicts $E^C \cap \bar{\Delta} \neq \emptyset$ in (5) and there exists no E satisfying the requirement. □

Based on the above preparations we now consider the following problem. Let two records i, j and a set of marginals $\mathcal{D} = \{D_1, \ldots, D_r\}$ be given. We are asked to find E such that E-swapping fixes all marginals of $\mathcal{D}$ and results in a different set of records. We consider this problem in terms of a graphical model. In the previous section we introduced a graph $G^{\mathcal{D}}$ generated by $\mathcal{D}$. Let $G_{\bar{\Delta}}$ denote the induced subgraph of $G^{\mathcal{D}}$ where the vertex set is restricted to $\bar{\Delta}$. Note that $G_{\bar{\Delta}}$ is a graph with the vertex set $\bar{\Delta}$ and an edge between $s, t \in \bar{\Delta}$ if and only if there exists $D \in \mathcal{D}$ such that $\{s, t\} \subset D$.

Recall that the variables s and t belonging to some $D \in \mathcal{D}$ either have to be swapped out simultaneously or stay together. It follows that any variable in a connected component of $G_{\bar{\Delta}}$ has to be swapped out simultaneously or stay together simultaneously. Therefore we have the following theorem, which is the main theorem of this paper.

Theorem 3.1. *Given two records i, j and a generating class $\mathcal{D}$, we can find $E \subset \Delta$ such that E-swapping is effective and fixes all D-marginals, $\forall D \in \mathcal{D}$, if and only if $G_{\bar{\Delta}}$ is not connected.*

Proof. As mentioned above, there exists no $E \subset \Delta$ such that E-swapping is effective and fixes all D-marginals in the case where $G_{\bar{\Delta}}$ is connected.

Conversely assume that $G_{\bar{\Delta}}$ is not connected. Let $\gamma_{\bar{\Delta}}$ be a connected component of $G_{\bar{\Delta}}$. Then for any two vertices $\{s, t\}$ such that $s \in \gamma_{\bar{\Delta}}$ and $t \in \bar{\Delta} \setminus \gamma_{\bar{\Delta}}$ there exists no $D \in \mathcal{D}$ satisfying $\{s, t\} \subset D$. Therefore if we set $E = \gamma_{\bar{\Delta}}$, E-swapping is effective and fixes all D-marginals. □

For example let $\mathcal{D}$ consists of all two-element sets of Δ. This $\mathcal{D}$ corresponds to the hierarchical model containing all two-variable interaction terms but not containing any higher order interactions terms. For this $\mathcal{D}$, $G^{\mathcal{D}}$ is the complete graph, corresponding to the saturated model.

If $\mathcal{D}$ consists of all two-element sets of Δ, i.e., if we have to fix all two-dimensional marginals, then $G^{\mathcal{D}}$ is complete and $G_{\bar{\Delta}}$ is also complete. In particular $G_{\bar{\Delta}}$ is connected and Theorem 3.1 says that we can not find an effective swapping fixing all two-dimensional marginals.

Let $\mathcal{S}^{\mathcal{D}}$ be the set of the minimal vertex separators of $G^{\mathcal{D}}$. It is well known that any $S \in \mathcal{S}^{\mathcal{D}}$ induces complete subgraph of $G^{\mathcal{D}}$ when $G^{\mathcal{D}}$ is chordal, that is, $\mathcal{D}$ is a generating class of a decomposable model. Denote the induced subgraph of $G^{\mathcal{D}}$ to $\Delta \setminus S$ by $G^{\mathcal{D}}_{\Delta\setminus S}$. Let $\mathrm{adj}(\alpha)$, $\alpha \in \Delta$ denote the set of vertices which are adjacent to α. Define $\mathrm{adj}(A)$ for $A \subset \Delta$ by $\mathrm{adj}(A) = \bigcup_{\delta\in A} \mathrm{adj}(\delta) \setminus A$. Then we obtain the following lemma.

Lemma 3.3. *$G_{\bar{\Delta}}$ is not connected if and only if there exist $S \in \mathcal{S}^{\mathcal{D}}$ and two connected components γ_α and γ_β of $G^{\mathcal{D}}_{\Delta\setminus S}$ such that*

$$S \cap \bar{\Delta} = \emptyset, \quad \gamma_\alpha \cap \bar{\Delta} \neq \emptyset, \quad \gamma_\beta \cap \bar{\Delta} \neq \emptyset. \tag{6}$$

Proof. Assume that $G_{\bar{\Delta}}$ is not connected. Let $\gamma_{\bar{\Delta},1}$ and $\gamma_{\bar{\Delta},2}$ be any two connected components of $G_{\bar{\Delta}}$. For any pair of vertices (α, β) such that $\alpha \in \gamma_{\bar{\Delta},1}$ and $\beta \in \gamma_{\bar{\Delta},2}$, $\mathrm{adj}(\gamma_{\bar{\Delta},1})$ is a (α, β)-separator (not necessarily minimal) in $G^{\mathcal{D}}$. Hence there exists $S_{\alpha,\beta} \in \mathcal{S}^{\mathcal{D}}$ such that $S_{\alpha,\beta} \subset \mathrm{adj}(\gamma_{\bar{\Delta},1})$. If there does not exist $S_{\alpha,\beta} \in \mathcal{S}$ satisfying $S_{\alpha,\beta} \cap \bar{\Delta} = \emptyset$, then $\mathrm{adj}(\gamma_{\bar{\Delta},1}) \cap \bar{\Delta} \neq \emptyset$, which contradicts that the intersections of $\gamma_{\bar{\Delta},1}$ and other connected components of $G_{\bar{\Delta}}$ are empty. Therefore there exists a minimal (α, β)-separator such that $S_{\alpha,\beta} \cap \bar{\Delta} = \emptyset$.

Since each of $\gamma_{\bar{\Delta},1}$ and $\gamma_{\bar{\Delta},2}$ is a connected component, $S_{\alpha,\beta}$ satisfying $S_{\alpha,\beta} \cap \bar{\Delta} = \emptyset$ also separates any pair of vertices in $\gamma_{\bar{\Delta},1}$ and $\gamma_{\bar{\Delta},2}$ other than (α, β). Hence $S_{\alpha,\beta}$ separates $\gamma_{\bar{\Delta},1}$ and $\gamma_{\bar{\Delta},2}$ in $G^{\mathcal{D}}$. This implies that $\gamma_{\bar{\Delta},1}$ and $\gamma_{\bar{\Delta},2}$ belong to different connected components of $G^{\mathcal{D}}_{\Delta\setminus S_{\alpha,\beta}}$. Therefore (6) is satisfied.

On the other hand if there exist S, γ_α and γ_β satisfying (6), it is obvious that $G_{\bar{\Delta}}$ is not connected. □

By the above lemma, we have the following corollary.

Corollary 3.1. *Given two records i, j and a generating class $\mathcal{D}$, we can find $E \subset \Delta$ such that E-swapping is effective and fixes all D-marginals, $\forall D \in \mathcal{D}$, if and only if there exist $S \in \mathcal{S}^{\mathcal{D}}$ and two connected components γ_α and γ_β of $G^{\mathcal{D}}_{\Delta\setminus S}$ satisfying (6), that is,*

$$i_S = j_S, \quad i_{\gamma_\alpha} \neq j_{\gamma_\alpha}, \quad i_{\gamma_\beta} \neq j_{\gamma_\beta}. \tag{7}$$

Theorem 3.1 and Corollary 3.1 are applicable to general hierarchical models. If $\mathcal{D}$ is a generating class of a graphical model associated with a graph G, then by definition $G^{\mathcal{D}} = G$. Therefore we have the following corollary concerning a graphical model.

Corollary 3.2. *Let $\mathcal{D}$ be a generating class of a graphical model associated with a graph G. For two records i, j define $\bar{\Delta}$ by (4). We can find $E \subset \Delta$ such that E-swapping of i and j is effective and fixes all D-marginals, $\forall D \in \mathcal{D}$, if and only if there exist $S \in \mathcal{S}^{\mathcal{D}}$ and two connected components γ_α and γ_β of $G_{\Delta\setminus S}$ satisfying (6), that is,*

$$i_S = j_S, \quad i_{\gamma_\alpha} \neq j_{\gamma_\alpha}, \quad i_{\gamma_\beta} \neq j_{\gamma_\beta}.$$

3.2 Searching another record for swapping

So far we have considered some necessary and sufficient conditions on E-swapping between two records i, j to be effective and fix D-marginals for general hierarchical models. In this section we consider to find another record which is swappable for a particular sample unique record i by using the results in the previous section.

Given a particular record i, by Corollary 3.1, we could scan through the microdata set for another record j satisfying the conditions of Corollary 3.1. Instead of checking the conditions Corollary 3.1 for each j, we could first construct the list $\mathcal{S}^{\mathcal{D}}$ of minimal vertex separators S and the connected components γ_α, γ_β of $G^{\mathcal{D}}_{\Delta\setminus S}$. The for a particular triple $(S, \gamma_\alpha, \gamma_\beta)$ we could check whether there exists another record j satisfying (7) of Corollary 3.1. Actually it is straightforward to check the existence of j satisfying (7). Since we require $i_S = j_S$, we only need to look at the slice of the contingency table given the value of i_S. Then in this slice we look at $\{i_{\gamma_\alpha}, i_{\gamma_\beta}\}$-marginal table. By the requirement $i_{\gamma_\alpha} \neq j_{\gamma_\alpha}, i_{\gamma_\beta} \neq j_{\gamma_\beta}$, we omit the "row" i_{γ_α} and the "column" i_{γ_β} from the marginal table. If the resulting table is non-empty, then we can find another record j in a diagonal position to i and we can swap observations in j and i. See Figure 1.

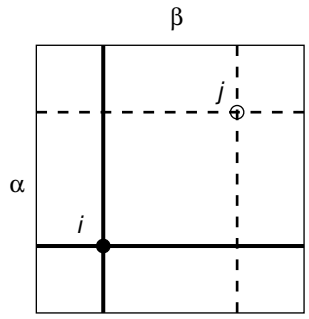

Figure 1: j swappable with i in a diagonal position

More precisely, for γ_α, γ_β, S, write $\gamma_{\alpha,\beta} = \gamma_\alpha \cup \gamma_\beta \cup S$. Define the subtable $\bar{\mathbf{n}}_{\gamma_{\alpha,\beta}}(i'_{\gamma_{\alpha,\beta}} \mid i_{\gamma_{\alpha,\beta}})$ by

$$\bar{\mathbf{n}}_{\gamma_{\alpha,\beta}}(i'_{\gamma_{\alpha,\beta}} \mid i_{\gamma_{\alpha,\beta}}) = \Big\{\bar{n}_{\gamma_{\alpha,\beta}}(i'_{\gamma_{\alpha,\beta}} \mid i_{\gamma_{\alpha,\beta}})\Big\} = \Big\{n_{\gamma_{\alpha,\beta}}(i'_{\gamma_{\alpha,\beta}}) \mid i'_{\gamma_\alpha} \neq i_{\gamma_\alpha}, i'_{\gamma_\beta} \neq i_{\gamma_\beta}, i'_S = i_S\Big\}.$$

Let $\bar{\mathbf{n}}_{\gamma_{\alpha,\beta}}(i'_{\gamma_{\alpha,\beta}} \mid i_{\gamma_{\alpha,\beta}}) \neq \mathbf{0}$ denote that there exists at least one positive count in $\bar{\mathbf{n}}_{\gamma_{\alpha,\beta}}(i'_{\gamma_{\alpha,\beta}} \mid i_{\gamma_{\alpha,\beta}})$. Then we have the following lemma. Proof is obvious and omitted.

Lemma 3.4. *There exists a record j with $i_S = j_S$, $i_{\gamma_\alpha} \neq j_{\gamma_\alpha}$, and $i_{\gamma_\beta} \neq j_{\gamma_\beta}$ if and only if $\bar{\mathbf{n}}_{\gamma_{\alpha,\beta}}(i'_{\gamma_{\alpha,\beta}} \mid i_{\gamma_{\alpha,\beta}}) \neq \mathbf{0}$.*

Lemma 3.4 is easy to check. Therefore it remains to compute the set of minimal vertex separators $\mathcal{S}^{\mathcal{D}}$ and the connected components of $G^{\mathcal{D}}_{\Delta\setminus S}$. Shiloach and Vishkin [1982] proposed an algorithm for computing connected components of a graph. On listing

minimal vertex separators there exist algorithms by Berry et al. [2000] and Kloks and Kratsch [1998]. The input of their algorithms is $G^{\mathcal{D}}$. However in our case generating class $\mathcal{D}$ is given in advance. It may be possible to obtain more efficient algorithms if we also use the information of $\mathcal{D}$ as the input.

The following algorithm searches another record j which is swappable for a sample unique record i and swaps them if it exists.

Algorithm 3.1 (Finding j swappable for i and swapping between i and j).
Input : $\mathbf{n}$, $\mathcal{D}$, $\mathcal{S}^{\mathcal{D}}$, i
Output : a post-swapped table $\mathbf{n}' = \{n'(i)\}$
begin
  $\mathbf{n}' \leftarrow \mathbf{n}$;
  for every $S \in \mathcal{S}^{\mathcal{D}}$ **do**
  begin
    compute connected components of $G^{\mathcal{D}}_{\Delta \setminus S}$;
    for every pair of connected components $(\gamma_\alpha, \gamma_\beta)$ **do**
    begin
      if $\bar{\mathbf{n}}_{\gamma_{\alpha,\beta}}(i'_{\gamma_{\alpha,\beta}} \mid i_{\gamma_{\alpha,\beta}}) \neq \mathbf{0}$ **then**
      begin
        select a marginal cell $i'_{\gamma_{\alpha,\beta}}$ such that $\bar{n}_{\gamma_{\alpha,\beta}}(i'_{\gamma_{\alpha,\beta}} \mid i_{\gamma_{\alpha,\beta}}) \neq 0$;
        select a cell $j \in \mathcal{I}$ such that $j_{\gamma_{\alpha,\beta}} = i'_{\gamma_{\alpha,\beta}}$;
        $E \leftarrow \gamma_\alpha$;
        E-swapping between i and j;
        $n'(i) \leftarrow n(i) - 1$;
        $n'(j) \leftarrow n(j) - 1$;
        $n'(j_E, i_{E^c}) \leftarrow n(j_E, i_{E^c}) + 1$;
        $n'(i_E, j_{E^c}) \leftarrow n(i_E, j_{E^c}) + 1$;
        exit ;
      end if
    end for
  end for
  if $\mathbf{n}' = \mathbf{n}$ **then** i is not swappable ;
end

In Takemura and Endo [2006] we applied this algorithm to a microdata set of $n = 9809$ records and $k = 8$ variables. There were 2243 sample unique records. We fitted a decomposable model to the 8-way contingency table to identify 50 risky records among the 2243 sample unique records. We then applied Algorithm 3.1 to check whether these 50 records are swappable or not. For most of these 50 records, Algorithm 3.1 quickly found another record for swapping. Therefore we found that Algorithm 3.1 is very practical in actual disclosure control procedures.

4 Some discussions

In this paper we considered swapping among two records. As mentioned above, if all two-dimensional marginals are fixed, then we can not swap among two records without disturbing some marginal. However when we consider swapping among more than two records, there are cases where we can fix all two-dimensional marginals, as illustrated by the following example. Consider a table of 4 records with 3 variables. Each variable has two levels (1 or 2).

x_1	x_2	x_3
1	1	1
1	2	2
2	2	1
2	1	2

In this example there is exactly 1 frequency for each 2-marginal. If we now circularly rotate the observations of x_3, we obtain the following table.

x_1	x_2	x_3
1	1	2
1	2	1
2	2	2
2	1	1

Then all 4 records are changed but all two-dimensional marginals are preserved. In fact this example correspond to a basic move of degree 4 (Diaconis and Sturmfels [1998]) of the Markov basis for $2 \times 2 \times 2$ contingency tables with fixed two-dimensional marginals. More complicated examples can be given by translating the moves of $3 \times 3 \times K$ tables of Aoki and Takemura [2003].

Dobra [2003] proved that there exists a Markov basis consisting of primitive moves for decomposable models. This implies the following fact in the case of decomposable models. If a particular record can be changed by swaps possibly involving more than 2 records, then it is always possible to change the record by a swap involving the record and another single record.

On the other hand Geiger et al. [2006] have shown that that primitive moves do not form a Markov basis for non-decomposable models. This implies that for non-decomposable models, there is a possibility of swapping of a sample unique record involving more than 2 records, even if it can not be swapped with another single record that can be checked by Algorithm 3.1 of Section 3.2.

The theory of Markov basis is concerned with the swappability of all records with arbitrary marginal counts. The investigation of this paper just asks whether a particular sample unique record can be swapped with other records in a particular data set. Therefore the problem considered here should be much easier than the problem of construction of Markov bases for general hierarchical models of contingency tables. Still it is not clear

at this point how to construct a practical algorithm for checking swappability of a particular record involving other two records, other three records etc. This problem is left for our future research.

A Equivalence of a primitive move and swapping of two records

An effective E-swapping (1) changes the cell frequencies of i, j, i', j' into

$$n(i) \to n(i) - 1, \quad n(j) \to n(j) - 1, \quad n(i') \to n(i') + 1, \quad n(j') \to n(j') + 1. \tag{8}$$

Hence the difference between the post-swapped and the pre-swapped tables is a primitive move. If E-swapping fixes all $\mathcal{D}$-marginals, the corresponding primitive move also fixes them.

Next we consider to show that any primitive move (8) for $\mathcal{D}$ can be expressed by E-swapping (1) for some $E \subset \Delta$. Write

$$i = (i_1, \ldots, i_k),\ j = (j_1, \ldots, j_k), \quad i' = (i'_1, \ldots, i'_k),\ j = (j'_1, \ldots, j'_k).$$

We first show that $\{i_m, j_m\} = \{i'_m, j'_m\}$ for $1 \le m \le k$. Since $\bigcup_t D_t = \Delta$, there exists t for any m such that m belongs to D_t. In the case where $i_{D_t} = j_{D_t}$, two records of $n_{D_t}(i_{D_t})$ have to be preserved in i'_{D_t} and j'_{D_t}. Hence $i'_{D_t} = j'_{D_t} = i_{D_t} = j_{D_t}$. On the other hand if $i_{D_t} \neq j_{D_t}$, each one record of both $n_{D_t}(i_{D_t})$ and $n_{D_t}(j_{D_t})$ have to be preserved in $\{i'_{D_t}, j'_{D_t}\}$, which implies $\{i_{D_t}, j_{D_t}\} = \{i'_{D_t}, j'_{D_t}\}$. Therefore we have $\{i_m, j_m\} = \{i'_m, j'_m\}$ for $1 \le m \le k$.

If we set

$$E = \{m \mid i'_m = j_m\} = \{m \mid i_m = j'_m\},$$

E satisfies (1). This completes the proof of the equivalence of E-swapping and primitive move for $\mathcal{D}$.

References

Satoshi Aoki and Akimichi Takemura. Minimal basis for a connected Markov chain over $3 \times 3 \times K$ contingency tables with fixed two-dimensional marginals. *Aust. N. Z. J. Stat.*, 45(2):229–249, 2003. ISSN 1369-1473.

Anne Berry, Jean Paul Bordat, and Olivier Cogis. Generating all the minimal separators of a graph. *Int. J. Found. Comput. Sci.*, 11(3):397–403, 2000.

Tore Dalenius and Steven P. Reiss. Data-swapping: a technique for disclosure control. *J. Statist. Plann. Inference*, 6(1):73–85, 1982. ISSN 0378-3758.

Persi Diaconis and Bernd Sturmfels. Algebraic algorithms for sampling from conditional distributions. *Ann. Statist.*, 26(1):363–397, 1998. ISSN 0090-5364.

Adrian Dobra. Markov bases for decomposable graphical models. *Bernoulli*, 9(6):1093–1108, 2003. ISSN 1350-7265.

Adrian Dobra and Seth Sullivant. A divide-and-conquer algorithm for generating Markov bases of multi-way tables. *Comput. Statist.*, 19(3):347–366, 2004. ISSN 0943-4062.

Dan Geiger, Chris Meek, and Bernd Sturmfels. On the toric algebra of graphical models. *Ann. Statist.*, 2006. To appear.

Daniel A. Klain and Gian-Carlo Rota. *Introduction to geometric probability*. Lezioni Lincee. [Lincei Lectures]. Cambridge University Press, Cambridge, 1997. ISBN 0-521-59362-X; 0-521-59654-8.

D. Kloks and D. Kratsch. Listing all minimal separators of a graph. *SIAM J. Comput.*, 27(3):605–613, 1998.

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

Jan Schlörer. Security of statistical databases: multidimensional transformation. *ACM Trans. Database Systems*, 6(1):95–112, 1981. ISSN 0362-5915.

Yossi Shiloach and Uzi Vishkin. An $O(\log n)$ parallel connectivity algorithm. *J. Algorithms.*, 3:57–67, 1982.

Akimichi Takemura. Local recording and record swapping by maximum weight matching for 0 disclosure control of microdata sets. *Journal of Official Statistics*, 18(2):275–289, 2002.

Akimichi Takemura and Yushi Endo. Evaluation of per-record identification risk and swappability of records in a microdata set via decomposable models. 2006. Technical Report METR 2006-17, University of Tokyo.

Leon Willenborg and Ton de Waal. *Elements of statistical disclosure control*, volume 155 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 2001. ISBN 0-387-95121-0.