ISTITUTO DI ANALISI DEI SISTEMI ED INFORMATICA
"Antonio Ruberti"
CONSIGLIO NAZIONALE DELLE RICERCHE

F. M. Malvestuto, E. Pourabbas

LOCAL COMPUTATION OF ANSWERS TO TABLE QUERIES ON SUMMARY DATABASES

R. 622 Novembre 2004

Francesco M. Malvestuto – Dipartiemnto di Informatica dell'Universtiá degli Studi di Roma "La Sapienza", via Salaria 12 - 00185 Roma, Italy. Email: malvestuto@di.uniroma1.it.

Elaheh Pourabbas – Istituto di Analisi dei Sistemi ed Informatica "Antonio Ruberti" del CNR, viale Manzoni 30 - 00185 Roma, Italy. Email: pourabbas@iasi.cnr.it.

ISSN: 1128–3378

Collana dei Rapporti dell'Istituto di Analisi dei Sistemi ed Informatica "Antonio Ruberti", CNR

viale Manzoni 30, 00185 ROMA, Italy

tel. ++39-06-77161
fax ++39-06-7716461
email: iasi@iasi.rm.cnr.it
URL: http://www.iasi.rm.cnr.it

Abstract

We address the problem of evaluating table queries from a summary database formed by a collection of pre-computed tables on certain measure variables. We assume that every table query asks for the distribution of a measure variable of interest, and that the summary database contains tables on the variable of interest as well as on other measure variables. If the requested distribution is none of the base tables and cannot be exactly derivable from none of them, then the answer to the query will be the result of an estimation procedure, which may bring up another measure variable that is correlated to the measure variable of interest. We give an estimation procedure that combines the "divide-and-conquer" principle with tree computations.

Keywords: Acyclic Hypergraph, Join Tree, Graham Reduction, Iterative Proportional Fitting Procedure, Maximum-Entropy Table, Minimum-Cross-Entropy Table

1. Introduction

A recent querying paradigm, called *On-Line Analytical Processing* (OLAP) [6], often involves complex queries over very large multidimensional relations or datacubes with category (or dimensional) and measure (or summary) attributes. Obtaining the exact answer to an OLAP query can be prohibitively expensive in terms of time and/or storage space in data warehouse environment. In order to reduce the computational effort, a promising approach is to store some aggregate data (as " materialized views") in a summary database, which is used to answer OLAP queries.

In this paper, we consider the problem of evaluating table queries from a summary database formed by a collection of pre-computed tables on certain measure variables. We suppose that each table query asks for the distribution of a measure variable of interest, called the *target variable* (e.g., Personnel 2001), by a set of category attributes (e.g., *gender*, *state*,...), and that the summary database contains tables on the target variable as well as on other measure variables. If the table requested by a query is none of the base tables and cannot be exactly derived from them [13] [5], then the query can be answered in an approximate (and, hopefully, accurate) way using some information-theoretic estimation criterion which, on demand of the user, may bring up an additional measure variable (e.g., Personnel 2000, Total-Income 2001), called the *auxiliary variable*. We can use as estimation criterion the principle of *minimum cross-entropy* or the principle of *maximum entropy* [17], depending on whether or not the target and auxiliary variables can be viewed as terms of a time series (see the following example). Both are based on the proportionality principle and have been successfully applied to the *small-area estimation* [9] and to the analysis of inter-industry transactions with *input-output matrices* [1] [10][18].

Example 1 A summary database contains four tables: two on the measure variable *Personnel 2001*, one table on the measure variable *Total-Income 2001*, and one table on the measure variable *Personnel 2000*. The two tables on *Personnel 2001* report the distributions $p_1(g, a)$ and $p_2(d, a)$ of employees in 2001 by *gender* and *age-class*, and by *department* and *age-class*, respectively. The table on *Total-Income 2001* reports the distribution $t(g, l)$ of total income in 2001 by *gender* and *level*. The table on *Personnel 2000* reports the distribution $q(g, d, a, l)$ of employees in 2000 by *gender*, *department*, *age-class* and *level*.
Suppose that a user asks for the distribution of employees in 2001 by *age-class* and *level*. Then, the query system advices the user that he will receive an estimate of the requested table and that he may tune the answer to some auxiliary variable. We now discuss three typical cases:

Case 1: the user selects no auxiliary variable. Let $\hat{p}(g, d, a)$ be the maximum entropy extension of $p_1(g, a)$ and $p_2(d, a)$. Then, the query will be answered by issuing the marginal on a and l of the distribution $\dfrac{\hat{p}(g, d, a)}{L}$, where L is the number of possible levels.

Case 2: the user selects *Total-Income 2001* as auxiliary variable. Let $\hat{p}(g, d, a)$ be as above and let $t(g)$ be the marginal on g of $t(g, l)$. Then, the query will be answered by issuing the marginal on a and l of the distribution $\dfrac{\hat{p}(g, d, a)t(g, l)}{t(g)}$.

Case 3: the user selects *Personnel 2000* as auxiliary variable. Let $\tilde{p}(g, d, a)$ be the minimum cross-entropy extension of $p_1(g, a)$ and $p_2(d, a)$ relative to the marginal $q(g, d, a)$ of the distri-

bution $q(g, d, a, l)$. Then, the query will be answered by issuing the marginal on a and l of the distribution $\dfrac{\tilde{p}(g, d, a)q(g, d, a, l)}{q(g, d, a)}$.

In this paper we show how to solve the problem of answering a table query using only tables stored in a summary database, referred to as the *Table-Query Problem*. The proposed procedure is inspired by the "divide-and-conquer" principle and generalizes that given in [17], which applies only to the query that asks for the distribution of the target variable by all the category attributes of the target tables and the auxiliary table.

The paper is structured as follows. In the next section, we state the two Proportional Estimation Models PEM1 and PEM2 and the Table-Query Problem. In Sections 3 and 4, we solve the Table-Query Problem under the models PEM1 and PEM2, respectively. Finally, Section 5 concludes.

2. The Table-Query Problem

Henceforth, we assume that all measure variables are of nonnegative-real type and of additive nature. Let X be a set of (category) attributes. The *domain* of X, written $dom(X)$, is the set of all semantically possible tuples on X; by $size(X)$ we denote the cardinality of $dom(X)$. Let $p(x)$ be a nonnegative real-valued function defined on $dom(X)$; the *support* of $p(x)$ is the relation with scheme X containing all tuples x with $p(x) \neq 0$. The pair $T = \langle X, p(x)\rangle$ defines a *(summary) table*, of which X is the *scheme* and $p(x)$ the *distribution*. Without loss of generality, we always assume that the data reported in every table are normalized to one. Let Y be a subset of X; the *marginal* of T with respect to Y is the table $T(Y) = \langle Y, p(y)\rangle$ where $p(y)$ is the marginal of the distribution $p(x)$, that is, $p(y) = \sum_x p(x)$ the summation being extended over all tuples x in $dom(X)$ whose restrictions to Y coincide with y. Note that the support of $p(y)$ is the (relation-theoretic) projection onto Y of the support of $p(x)$. We also admit the case Y is empty; then, $p(y)$ is the unity. Let $\mathcal{T} = \{T_1, ..., T_n\}$ be a set of tables, where T_i has scheme X_i, for all i. The set X given by the union of the schemes X_i of the tables T_i and their collection $\mathbf{H}$ will be referred to as the *base set* and the *scheme* of $\mathcal{T}$, respectively. The table set $\mathcal{T}$ is *consistent* if there exists at least one table T with scheme X such that the marginal of T with respect to X_i coincides with T_i, for all i. Such a table is called a *universal table* of $\mathcal{T}$.

Suppose we are given a table query and that the user has selected a certain auxiliary variable. Let us assume that the summary database contains the set of tables $\mathcal{T} = \{T_1, ..., T_n\}$ on the target variable, where T_i has scheme X_i, and the table $\langle Y, q(y)\rangle$ on the auxiliary variable. Consider the following two Proportional Estimation Models where $p(x, y)$ denotes the distribution of an unknown table with scheme $X \cup Y$.

PEM 1
Marginal constraints: $p(x_i) = p_i(x_i), \quad i = 1, \ldots, n$ *Proportionality criterion:* Let $Z = X \cap Y$. There exist real-valued functions $g_1(x_1), \ldots, g_n(x_n)$ such that the factorization $p(x, y) = g_1(x_1) \cdots g_n(x_n)\dfrac{q(y)}{q(z)}$ holds for every tuple (x, y) in the support of $p(x, y)$.

PEM 2
Marginal constraints: $p(x_i) = p_i(x_i), \quad i = 1, \ldots, n$ *Proportionality criterion:* There exist real-valued functions $g_1(x_1), \ldots, g_n(x_n)$ such that the factorization $p(x, y) = g_1(x_1) \cdots g_n(x_n) q(y)$ holds for every tuple (x, y) in the support of $p(x, y)$.

Using the results proven in our previous paper [17], one has that: PEM1 has a unique solution, we denote by $\hat{p}(x, y)$, and PEM2 has a solution if and only if there exists a universal table $T = \langle X, p(x) \rangle$ of $\mathcal{T}$ such that the support of the marginal of $p(x)$ with respect to $Z = X \cap Y$ is contained in the support of the marginal of $q(y)$ with respect to Z, and if this is the case then PEM2 has a unique solution, we denote by $\tilde{p}(x, y)$. At this point, we can state the *Table-Query Problem* we want to solve:

Given a nonempty subset U of $X \cup Y$, find the marginal with respect to U of $\hat{p}(x, y)$ or $\tilde{p}(x, y)$ depending on whether PEM1 or PEM2 is in use. In [17] the Table-Query Problem was solved for the special case that $U = X \cup Y$. We now state some useful formulas for solving the Table-Query Problem in the general case. By summing out the variables in $Y - Z$ in the functional expressions of $\hat{p}(x, y)$ and $\tilde{p}(x, y)$, we obtain

$$\hat{p}(x) = g_1(x_1) \ldots g_n(x_n) \tag{1}$$

for every tuple x in the support of $\hat{p}(x)$, and

$$\tilde{p}(x) = g_1(x_1) \ldots g_n(x_n) q(z) \tag{2}$$

for every tuple x in the support of $\tilde{p}(x)$. Formulas (1) and (2) lead to the following expressions for the solutions to PEM1 and PEM2:

$$\hat{p}(x, y) = \hat{p}(x) \frac{q(y)}{q(z)} \qquad \tilde{p}(x, y) = \tilde{p}(x) \frac{q(y)}{q(z)} \tag{3}$$

Using formulas (3), we can easily find the expressions for $\hat{p}(u)$ and $\tilde{p}(u)$. Let $X' = (X - Y) \cap U$, $Y' = (Y - X) \cap U$, $Z' = Z \cap U$ and $Z'' = Z - U$. Then, $U = X' \cup Y' \cup Z'$ and we have:

$$\begin{aligned} \hat{p}(u) &= \textstyle\sum_{z''} \dfrac{\hat{p}(x', z', z'') q(y', z', z'')}{q(z', z'')} \quad \text{(i)} \\ \tilde{p}(u) &= \textstyle\sum_{z''} \dfrac{\tilde{p}(x', z', z'') q(y', z', z'')}{q(z', z'')} \quad \text{(ii)} \end{aligned} \tag{4}$$

Example 2 Suppose that a user asks for the table on a certain measure variable with scheme *abdhk*. Let $\mathcal{T} = \{T_1, \ldots, T_{12}\}$ be the set of the base tables on the target variable (see Figure 1).

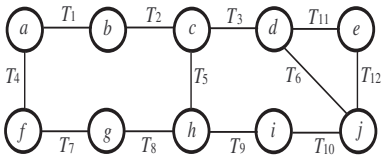

Figure 1: The table set $\mathcal{T}$ on the target variable

Note that the scheme $\mathbf{H} = \{ab, af, bc, cd, ch, de, dj, ej, fg, gh, hi, ij\}$ of $\mathcal{T}$ does not contain the attribute k. Suppose that the user selects an auxiliary variable for which there exists a base table with scheme $bcgkl$ and distribution q. Then, $4(i)$ and $4(ii)$ read:

$$\hat{p}(a,b,d,h,k) = \sum_{c,\,g} \frac{\hat{p}(a,b,c,d,g,h)q(b,c,g,k)}{q(b,c,g)}$$
$$\tilde{p}(a,b,d,h,k) = \sum_{c,\,g} \frac{\tilde{p}(a,b,c,d,g,h)q(b,c,g,k)}{q(b,c,g)}$$

Suppose that we are able to compute the distributions $\hat{p}(x', z', z'')$ and $\tilde{p}(x', z', z'')$. Then, the procedure below yields $\hat{p}(u)$ and $\tilde{p}(u)$.

MARGINAL
(1) Find the distribution $\hat{p}(x', z', z'')$ (respectively, and $\tilde{p}(x', z', z'')$). (2) Marginalize $q(y)$ with respect to $Y' \cup Z$ and Z. (3) Compute $\hat{p}(u)$ (respectively, $\tilde{p}(u)$) using $4(i)$ (respectively, $4(ii)$).

The execution of Steps 2 and 3 of MARGINAL is a matter of routine; so, we focus on Step 1. Let $V = X' \cup Z$. Of course, the distributions $\hat{p}(v)$ and $\tilde{p}(v)$ are also the marginals with respect to V of $\hat{p}(x)$ and $\tilde{p}(x)$, respectively. Therefore, Step 1 requires solving the following problem:

Find the marginal with respect to V of $\hat{p}(x)$ (or of $\tilde{p}(x)$). A brute-force approach to solving this problem consists in first finding $\hat{p}(x)$ (or $\tilde{p}(x)$) and, then, marginalizing it. We now show how to compute $\hat{p}(x)$ and $\tilde{p}(x)$. First of all, observe that they are the distributions of two universal tables of $\mathcal{T}$, we denote by $\hat{T}$ and $\tilde{T}$, both of which, by formulas (1) and (2), have the following form:

$$f_1(x_1) \dots f_n(x_n)\pi(x) \tag{5}$$

for some real-valued functions $f_1(x_1), \dots, f_n(x_n)$ and for some distribution $\pi(x)$ over X. Explicitly, $\pi(x) = \dfrac{1}{size(X)}$ for $\hat{p}(x)$, and $\pi(x) = \dfrac{q(z)}{size(X - Z)}$ for $\tilde{p}(x)$. Now, it is well-known [3] [7] that, if $p(x)$ is the distribution of a universal table of $\mathcal{T}$ having the form (5), then $p(x)$ can be computed using the iterative procedure, called *Iterative Proportional Fitting Procedure* (IPFP) [8], which starts with the *zero approximation* $p^{[0]}(x) = \pi(x)$ and determines the *higher-order approximations* to $p(x)$ according to the following computation scheme:

first iteration cycle	$p^{[1]}(x)$	$\dots$	$p^{[n]}(x)$
second iteration cycle	$p^{[n+1]}(x)$	$\dots$	$p^{[2n]}(x)$
$\dots$	$\dots$	$\dots$	$\dots$
h-th iteration cycle	$p^{[hn+1]}(x)$	$\dots$	$p^{[hn+n]}(x)$
$\dots$	$\dots$	$\dots$	$\dots$

where the approximation $p^{[hn+i]}(x)$ in the $(h+1)$-th iteration cycle, $1 \leq i \leq n$, is obtained by fitting the approximation $p^{[hn+i-1]}(x)$ to the distribution $p_i(x_i)$ of the base table T_i:

$$p^{[hn+i]}(x) = \frac{p_i(x_i)}{p^{[hn+i-1]}(x_i)} p^{[hn+i-1]}(x).$$

From an information-theoretic point of view, the distribution $p(x)$ minimizes the cross-entropy relative to $\pi(x)$ (see Section A of the Appendix for information-theoretic definitions). So, the distribution of the universal table $\hat{T}$ minimizes the cross-entropy relative to the distribution $\frac{1}{size(X)}$ or, equivalently, maximizes the entropy (see Section A of the Appendix), and the distribution of the universal table $\tilde{T}$ minimizes the cross-entropy relative to the distribution $\frac{q(z)}{size(X-Z)}$. Accordingly, $\hat{T}$ will be referred to as the *maximum entropy universal table* [11] [12] (the *ME universal table*, for short) of $\mathcal{T}$, and $\tilde{T}$ as the *minimum cross-entropy universal table* relative to $\frac{q(z)}{size(X-Z)}$ (the *q-mCE universal table*, for short) of $\mathcal{T}$. Efficient procedures for computing the distributions of $\hat{T}$ and $\tilde{T}$ can be found in [11] [12] [2] and in [17], respectively. Once $\hat{p}(x)$ (or $\tilde{p}(x)$) have been computed, its marginal with respect to V can be easily obtained. However, as is shown in Sections 3 and 4, in most cases both $\hat{p}(v)$ and $\tilde{p}(v)$ can be computed without passing through the computation of $\hat{p}(x)$ and $\tilde{p}(x)$.

3. Marginalizing $\hat{T}$ with respect to V

An efficient procedure for computing the distribution $\hat{p}(v)$ from $\mathcal{T}$ rests on the notion of "collapsibility" [14] we now recall. Let $\mathbf{H} = \{X_1, \ldots, X_n\}$ be the scheme of $\mathcal{T}$ and X its base set; the *projection* of $\mathcal{T}$ onto a subset W of X is the table set $\mathcal{T}(W) = \{T_1(X_1 \cap W), \ldots, T_n(X_n \cap W)\}$, where redundant tables are omitted. The ME universal table $\hat{T}$ of $\mathcal{T}$ is *collapsible* onto W if the marginal of $\hat{T}$ with respect to W coincides with the ME universal table of the projection of $\mathcal{T}$ onto W. So, if W is a (possibly improper) superset of V that the ME universal table of $\mathcal{T}$ is collapsible onto, then $\hat{p}(v)$ can be obtained by first computing the ME universal table $\hat{T}(W)$ of $\mathcal{T}(W)$ and, then, marginalizing the distribution of $\hat{T}(W)$ with respect to V. The best choice for W will fall upon a minimal superset of V that the ME universal table of $\mathcal{T}$ is collapsible onto. Such a superset of V is unique and is called the *closed hull* of V in $\mathbf{H}$ [14]; furthermore, it coincides with the "canonical closure" [15] [16] of $\mathbf{H}$ when $\mathbf{H}$ is viewed as a hypergraph (see Section B of the Appendix for hypergraph-theoretic definitions). The procedure for finding the closed hull of V in $\mathbf{H}$ is based on the notion of the *compaction* of $\mathbf{H}$ [15] [16], which is the finest of the acyclic covers $\mathbf{K}$ such that the ME universal table of $\mathcal{T}$ is collapsible onto each edge of $\mathbf{K}$. It has a number of nice properties, two of which read: (a) the separators of $\mathbf{H}$ and of the compaction of $\mathbf{H}$ are the same; (b) if $\mathbf{H}$ is acyclic, then (and only then) the compaction of $\mathbf{H}$ coincides with $\mathbf{H}$. Let $\mathbf{K}$ be the compaction of $\mathbf{H}$. For each edge C of $\mathbf{K}$, we call the table set $\mathcal{T}(C)$ a *component* of $\mathcal{T}$.

Example 2 (*continued*). The compaction of $\mathbf{H}$ is $\mathbf{K} = \{abcfgh, cdhij, dej\}$. The components of the table set of Figure 1 are shown in Figure 2.

Given $\mathbf{H}$, the compaction $\mathbf{K} = \{C_1, \ldots, C_m\}$ of $\mathbf{H}$ and the set V, the closed hull of V in $\mathbf{H}$, say W, can be determined using the CLOSED HULL algorithm [15], whose Step 1 performs the *selective reduction* of $\mathbf{K}$ with *sacred set* V [19].

It should be noted that the sets $E_1, \ldots, E_k$ of Step 2 of CLOSED HULL are exactly the edges of the subhypergraph $\mathbf{K}(W)$ of $\mathbf{K}$ induced by W and that $\mathbf{K}(W)$ is an acyclic hypergraph. Moreover, since by property (b) the compaction of $\mathbf{K}$ is $\mathbf{K}$ itself, the set W is the closed hull of itself not only in $\mathbf{H}$ but also in $\mathbf{K}$. Finally, if $\mathbf{H}$ is acyclic, then $\mathbf{H} = \mathbf{K}$ and $\mathbf{K}(W) = \mathbf{H}(W)$. After determining the closed hull W of V in $\mathbf{H}$, the distribution $\hat{p}(w)$ can be obtained without passing through the computation of $\hat{p}(x)$ simply by applying the IPFP to $\mathcal{T}(W)$ with zero approximation

8.

the distribution $\dfrac{1}{size(W)}$.

CLOSED HULL
(1) Repeatedly apply the following two operations until neither can be longer applied: (i) Delete a vertex of $\mathbf{K}$ if it is not in V and belongs to exactly one edge; (ii) Delete an edge of $\mathbf{K}$ if it is contained in another edge. (2) Let $\mathbf{K}' = \{C'_{j_1}, \ldots, C'_{j_k}\}$ be the resulting hypergraph, where C'_{j_h} is the "residual part" of the edge C_{j_h} of $\mathbf{K}$. For each h, $1 \leq h \leq k$, set E_h to C'_{j_h} if C'_{j_h} is contained in some edge X_i of $\mathbf{H}$, and to C_{j_h} otherwise. (3) Set W to the empty set. For each h, $1 \leq h \leq k$, set $W := W \cup E_h$.

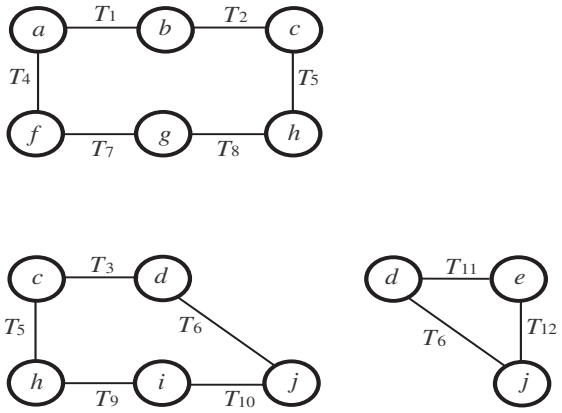

Figure 2: The components of the table set $\mathcal{T}$

Example 2 (*continued*). With $V = abcdgh$, the selective reduction of $\mathbf{K}$ with sacred set V (see Step 1 of CLOSED HULL) is the hypergraph $\mathbf{K}' = \{abcgh, cdh\}$, where $abcgh$ and cdh are the residual parts of the edges $abcfgh$ and $cdhij$ of $\mathbf{K}$, respectively. Since neither $abcgh$ nor cdh is contained in any edge of $\mathbf{H}$, the result of Step 3 of CLOSED HULL is $W = abcdfghij$, which hence is the closed hull of V. So, $\hat{p}(w)$ can be obtained as the ME universal table of the projection of $\mathcal{T}$ onto W (see Figure 3), that is, by applying the IPFP to $\mathcal{T}(W)$ with zero approximation the distribution $\dfrac{1}{size(abcdfghij)}$.

However, we can furthermore reduce the time and space costs using the implementation of the IPFP given in [11] [12] for computing the ME universal table $\hat{P}$ of any consistent table set $\mathcal{P}$ with scheme $\mathbf{R}$, we now recall. The implementation is based on the following two properties of $\hat{P}$ with respect to any acyclic cover $\mathbf{S}$ of $\mathbf{R}$:

(i) $\hat{P}$ coincides with the ME universal table of the set of the marginals of $\hat{P}$ with respect to the edges of $\mathbf{S}$, and

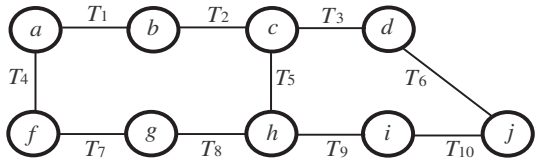

Figure 3: The projection of $\mathcal{T}$ onto W

(ii) the ME universal table of the set of the marginals of $\hat{P}$ with respect to the edges of $\mathbf{S}$ has a closed-form expression which, for any *join tree* J of $\mathbf{S}$ (see Section B of the Appendix), consists of a ratio whose numerator is the product of the marginals of $\hat{P}$ with respect to the labels of nodes of J (i.e., with respect to the edges of $\mathbf{S}$) and whose denominator is the product of the marginals of $\hat{P}$ with respect to the labels of arcs of J. Such a closed-form expression of $\hat{P}$ will be referred to as the *tree formula* generated by $\mathbf{S}$.

A first consequence is that, with $\mathcal{P} = \mathcal{T}$, $\mathbf{R} = \mathbf{H}$ and $\mathbf{S} = \mathbf{K}$, where $\mathbf{K}$ is the compaction of $\mathbf{H}$, the entries in the tree formula for $\hat{T}$ generated by $\mathbf{K}$ can be computed locally. More precisely, since $\hat{T}$ is collapsible onto each edge of $\mathbf{K}$, the marginal of $\hat{T}$ with respect to each edge of $\mathbf{K}$ can be computed as the ME universal table of the corresponding component of $\mathcal{T}$; moreover, by property (a) of $\mathbf{K}$, the separators of $\mathbf{K}$ are the same as $\mathbf{H}$, so that the marginal of $\hat{T}$ with respect to each separator of $\mathbf{K}$ can be obtained by marginalizing some table T_i in $\mathcal{T}$.

Example 2 (*continued*). Recall that the compaction of $\mathbf{H}$ is $\mathbf{K} = \{abcfgh, cdhij, dej\}$. The separators of $\mathbf{K}$ (and, hence, of $\mathbf{H}$) are ch and dj. A join tree of $\mathbf{K}$ is shown in Figure 4.

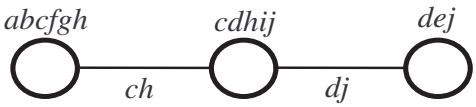

Figure 4: The join tree of $\mathbf{K}$

Therefore, the tree formula for $\hat{T}$ generated by $\mathbf{K}$ reads:

$$\frac{\hat{p}(abcfgh)\hat{p}(cdhij)\hat{p}(dej)}{p_5(ch)p_6(dj)}$$

where $\hat{p}(abcfgh)$, $\hat{p}(cdhij)$ and $\hat{p}(dej)$ can be computed as the distributions of the ME universal tables of the components $\mathcal{T}(abcfgh)$, $\mathcal{T}(cdhij)$ and $\mathcal{T}(dej)$ of $\mathcal{T}$ (see Figure 2), respectively, that is, by applying the IPFP procedure to $\mathcal{T}(abcfgh)$, $\mathcal{T}(cdhij)$ and $\mathcal{T}(dej)$ with zero approximations $\frac{1}{size(abcfgh)}$, $\frac{1}{size(cdhij)}$ and $\frac{1}{size(dej)}$, respectively.

We now apply the technique above to compute the ME universal table of the table set $\mathcal{T}(W)$ with scheme $\mathbf{H}(W)$. With $\mathcal{P} = \mathcal{T}(W)$, $\mathbf{R} = \mathbf{H}(W)$ and $\mathbf{S} = \mathbf{K}(W)$, the entries in the tree formula for $\hat{T}(W)$ generated by $\mathbf{K}(W)$ can be computed locally. Explicitly, with the notation of Closed Hull, the marginal of $\hat{T}(W)$ with respect to each edge E_h of $\mathbf{K}(W)$ can be computed as the ME universal table of the corresponding component $\mathcal{T}(E_h)$ of $\mathcal{T}(W)$, and the marginal of

$\hat{T}(W)$ with respect to each separator of $\mathbf{K}(W)$ can be computed as the marginal of some table in $\mathcal{T}(W)$.

Example 2 (*continued*). Recall that $W = abcdfghij$. The compaction of $\mathbf{H}(W)$ is $\mathbf{K}(W) = \{abcfgh, cdhij\}$. The separator of $\mathbf{K}(W)$ is ch. A join tree of $\mathbf{K}(W)$ is shown in Figure 5.

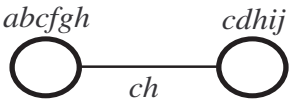

Figure 5: The join tree of $\mathbf{K}(W)$

Therefore, the tree formula for $\hat{T}(W)$ generated by $\mathbf{K}(W)$ reads:

$$\frac{\hat{p}(abcfgh)\hat{p}(cdhij)}{p_5(ch)}$$

where the distributions $\hat{p}(abcfgh)$ and $\hat{p}(cdhij)$ are computed as above.

Finally, once $\hat{T}(W)$ has been computed using its tree formula generated by $\mathbf{K}(W)$, $\hat{T}(V)$ can be obtained by marginalization. However, we can do better. Suppose that we have already found the entries in the tree formula for $\hat{T}(W)$ generated by $\mathbf{K}(W)$; at this point, instead of computing $\hat{T}(W)$, we soon marginalize the factors $\hat{p}(e_h)$ of the numerator, for each edge E_h of $\mathbf{K}(W)$, with respect to the edge C'_{j_h} of $\mathbf{K}'$. What we obtain is a tree formula generated by $\mathbf{K}'$, which provides a closed-form expression of $\hat{T}(W')$, where W' is the vertex set of $\mathbf{K}'$, that is, $W' = W - \bigcup_{h=1,\ldots,k}(E_h - C'_{j_h})$. Finally, after computing $\hat{T}(W')$, we, marginalize $\hat{T}(W')$ with respect to V.

Example 2 (*continued*). Recall that $\mathbf{K}' = \{abcgh, cdh\}$ and $\mathbf{K}(W) = \{abcfgh, cdhij\}$. After computing the distributions $\hat{p}(abcfgh)$ and $\hat{p}(cdhij)$, we soon marginalize them with respect to the edges $abcgh$ and cdh of $\mathbf{K}'$, respectively. The vertex set of of $\mathbf{K}'$ is $W' = abcdgh$ and the tree formula for $\hat{T}(W')$ generated $\mathbf{K}'$ reads:

$$\hat{p}(abcdgh) = \frac{\hat{p}(abcgh)\hat{p}(cdh)}{p_5(ch)}$$

4. Marginalizing $\tilde{T}$ with respect to V

The notion of collapsibility of the ME universal table of $\mathcal{T}$ naturally generalizes to the q-mCE universal table of $\mathcal{T}$ as follows. The table $\tilde{T}$ is *collapsible* onto W if the marginal of $\tilde{T}$ with respect to W coincides with the minimum cross-entropy universal table of $\mathcal{T}(W)$ relative to the marginal of $\dfrac{q(z)}{size(X-Z)}$ with respect to W. Unfortunately, at the present the uniqueness of a minimal superset of V that $\tilde{T}$ is collapsible onto is an open problem. Nevertheless, we can get an effective

procedure for computing the distribution $\tilde{p}(v)$ as follows. Given $\mathcal{T} = \{T_1, \ldots, T_n\}$ with scheme $\mathbf{H} = \{X_1, \ldots, X_n\}$, let us consider the (partially specified) table set $\mathcal{T}^* = \{\tilde{T}(Z), T_1, \ldots, T_n\}$, where redundant tables are omitted. It has scheme $\mathbf{H}^* = \{Z, X_1, \ldots, X_n\}$ where redundant edges are omitted. Then, it can be proven [17] that $\tilde{T}$ coincides with the ME universal table of $\mathcal{T}^*$. However, owing to the incompleteness of $\mathcal{T}^*$, we can seldom apply the technique developed in Section 3 to compute $\tilde{T}$ from $\mathcal{T}^*$. To see it, let $\mathbf{K}^*$ be the compaction of $\mathbf{H}^*$. Note that Z is a partial edge of $\mathbf{K}^*$, that is, Z is contained in at least one edge of $\mathbf{K}^*$. Of course, the closed hull W of V in $\mathbf{H}^*$ can be computed as in Section 3; but, like $\mathcal{T}^*$, also the projection of $\mathcal{T}^*$ onto W is partially specified unless the intersection of Z with each edge of $\mathbf{K}^*$ is contained in some edge X_i of $\mathbf{H}$. In what follows, we assume that this is not the case for, otherwise, $\tilde{T}$ does coincide with $\hat{T}$ [17].

Example 2 (*continued*). Recall that $Z = bcg$. The table set $\mathcal{T}^*$ is obtained from $\mathcal{T}$ by replacing the table T_2 by $\tilde{T}(Z)$ (see Figure 6).

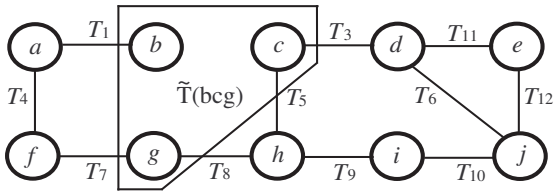

Figure 6: The table set $\mathcal{T}^*$

Analogously, the hypergraph $\mathbf{H}^*$ is obtained from $\mathbf{H}$ by replacing the edge $X_2 = bc$ by $Z = bcg$. The compaction of $\mathbf{H}^*$ is $\mathbf{K}^* = \{abfg, bcg, cgh, cdhij, dej\}$ and the components of $\mathcal{T}^*$ are shown in Figure 7.
With input $\mathbf{K}^* = \{abfg, bcg, cgh, cdhij, dej\}$ and $V = abcdgh$, CLOSED HULL yields $W = abcdfghij$. The projection of $\mathcal{T}^*$ onto W is shown in Figure 8.

In order to overcome the above-mentioned difficulty, we now introduce a suitable acyclic cover of $\mathbf{K}^*$. Let $\mathbf{K}^* = \{C_1, \ldots, C_m\}$ and let us assume that for some k, $1 \leq k \leq m-1$, $C_{k+1}, \ldots, C_m$ are the edges of $\mathbf{K}^*$ for which the set $Z \cap C_j$ is neither empty nor contained in any edge X_i of $\mathbf{H}$. Let $C = \bigcup_{j=k+1,\ldots,m} C_j$. It is easy to see that the hypergraph $\mathbf{K} = \{C, C_1, \ldots, C_k\}$ is an acyclic cover of $\mathbf{K}^*$, and each separator of $\mathbf{K}$ is contained in some edge X_i of $\mathbf{H}$. Moreover, by properties (i) and (ii) of acyclic schemes, one has that: $\tilde{T}$ coincides with the ME universal table of the set of the marginals of $\tilde{T}$ with respect to the edges of $\mathbf{K}$, and there is a tree formula for $\tilde{T}$ generated by $\mathbf{K}$. Finally, the marginal of $\tilde{T}$ with respect to each edge of $\mathbf{K}$ is the ME universal table of the corresponding projection of $\mathcal{T}^*$. Note that, for each j ($j = 1, \ldots, k$), the projections of $\mathcal{T}^*$ and $\mathcal{T}$ onto the edge C_j of $\mathbf{K}$ are the same (up to redundant tables) and, hence, their ME universal tables do coincide so that $\tilde{p}(C_j)$ can be computed by applying the IPFP procedure to $\mathcal{T}(C_j)$ with zero approximation $\dfrac{1}{size(C_j)}$. On the other hand, $\tilde{T}$ is collapsible onto the edge C of $\mathbf{K}$ [17] so that $\tilde{p}(C)$ can be computed by applying the IPFP procedure to $\mathcal{T}(C)$ with zero approximation $\dfrac{q(z)}{size(C - Z)}$.

12.

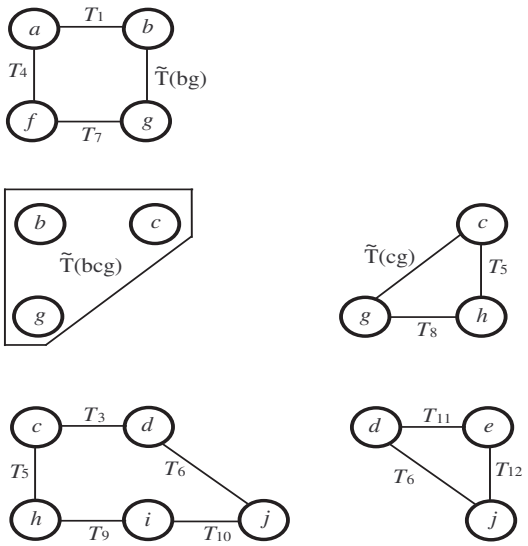

Figure 7: The components of $\mathcal{T}^*$

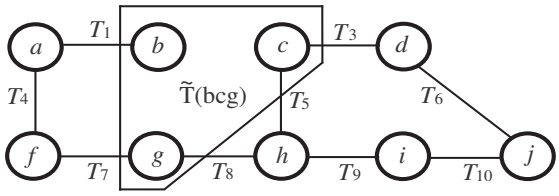

Figure 8: The projection on table set $\mathcal{T}^*$ onto W

Example 2 (*continued*). Recall that $Z = bcg$. The only edges C_j of $\mathbf{K}^*$ for which the set $Z \cap C_j$ is empty or is contained in some edge X_i of $\mathbf{H}$ are $cdhij$ and dej. Therefore, $C = abfg \cup bcg \cup cgh = abcfgh$ and $\mathbf{K} = \{abcfgh, cdhij, dej\}$. A join tree of $\mathbf{K}$ is shown in Figure 4 and the tree formula for $\tilde{T}$ generated $\mathbf{K}$ reads:

$$\frac{\tilde{p}(abcfgh)\tilde{p}(cdhij)\tilde{p}(dej)}{p_5(ch)p_6(dj)}$$

where the distribution $\tilde{p}(abcfgh)$ is computed by applying the IPFP procedure to $\mathcal{T}(abcfgh)$ with zero approximation $\dfrac{q(bcg)}{size(afh)}$, and the distributions $\tilde{p}(cdhij)$ and $\tilde{p}(dej)$ are computed in the same way as $\hat{p}(cdhij)$ and $\hat{p}(dej)$ (see above).

Turning to our problem of computing $\tilde{p}(v)$, it is sufficient to note that, since Z is contained in V, if we compute the closed hull of V in $\mathbf{K}$, it will be a superset of C. So, after determining the closed hull of V in $\mathbf{K}$, we can compute $\tilde{p}(v)$ using the marginalization technique employed for $\hat{p}(v)$ (see above).

Example 2 (*continued*). The closed hull of $V = abcdgh$ in $\mathbf{K}$ is $W = abcdghij$. The tree formula for $\tilde{T}(W)$ generated $\mathbf{K}(W)$ reads:

$$\frac{\tilde{p}(abcfgh)\tilde{p}(cdhij)}{p_5(ch)}$$

where the distributions $\tilde{p}(abcfgh)$ and $\tilde{p}(cdhij)$ are computed as above. Instead of computing $\tilde{p}(w)$ using the tree formula above, we soon marginalize $\tilde{p}(abcfgh)$ with respect to $abcgh$ and $\tilde{p}(cdhij)$ with respect to cdh. Thus, we obtain the tree formula for $\tilde{T}(abcdgh)$:

$$\frac{\tilde{p}(abcgh)\tilde{p}(cdh)}{p_5(ch)},$$

After computing the distribution of $\tilde{T}(abcdgh)$ using the tree formula above, we can finally obtain $\tilde{p}(v)$ by marginalization.

5. Conclusions

We have considered the problem of estimating the answer to a table query using a table set on the target variable and a table on an auxiliary variable selected by the user. We have shown that such a query can be answered with "local" computation, that is, using a (hopefully minimal) subset of the table set on the target variable and applying the principle of "divide-and-conquer". A direction for future research is the generalization of this approach to the case that also the information on the auxiliary variable is stored in a table set.

References

[1] M. Bacharach, *Biproportional Matrices and Input-Output Change*. Cambridge: University Press, 1970.

[2] J.-H. Badsberg and F. M. Malvestuto, "An implementation of the iterative proportional fitting procedure by propagation trees," *Computational Statistics and Data Analysis*, vol. 37, pp. 297–322, 2001.

[3] S. F. Bishop and P. Holland, *Discrete Multivariate Analysis*. MIT-Press, 1975.

[4] D. M. C. Beeri, R. Fagin and M. Yannakakis, "On the desirability of acyclic database schemes," *Journal of ACM*, vol. 30, pp. 479–513, 1983.

[5] H. V. J. C. Faloutsos and N. D. Sidiropoulos, "Recovering information from summary data," *Proceedings of the* 23*rd VLDB Conference*, pp. 36–45, 1997.

[6] S. Chauduri and U. Dayal, "An overview of data warehousing and olap technology," *ACM SIGMOD Record*, vol. 26, pp. 65–74, 1997.

[7] I. Csiszár, "*I*-divergence geometry of probability distributions and minimization problems," *The Annals of Probability*, vol. 3, pp. 146–158, 1975.

[8] W. E. Deming and F. F. Stephan, "On a least square adjustment of a sampled frequency table when the expected marginal totals are known," *Annals of Mathematical Statistics*, vol. 11, pp. 427–444, 1940.

[9] M. Ghosh and J. N. K. Rao, "Small area estimation: An appraisal," *Journal of Statistical Science*, vol. 9, pp. 55–93, 1994.

[10] W. W. Leontief and A. Strout, "Multiregional input-output analysis," *Structural Interdependence and Economic Development. T. Bama (Ed.)*, pp. 119–169, 1963.

[11] F. M. Malvestuto, "Answering queries in categorical databases," *Proc. of the 6th ACM Symp. on Principles of Database Systems*, pp. 87–96, 1987.

[12] F. M. Malvestuto, "A universal table model for categorical databases," *Information Sciences*, vol. 49, pp. 203–223, 1989.

[13] F. M. Malvestuto, "A universal - scheme approach to statistical databases containing homogeneous summary tables," *ACM Trans. on Database Systems*, vol. 18, pp. 678–708, 1993.

[14] F. M. Malvestuto, "A hypergraph-theoretic analysis of collapsibility and decomposability for extended log-linear models," *Statistics and Computing*, vol. 11, pp. 155–169, 2001.

[15] F. M. Malvestuto and M. Moscarini, "A fast algorithm for query optimization in universal-relation databases," *J. Computer and System Sciences*, vol. 56, pp. 299–309, 1998.

[16] F. M. Malvestuto and M. Moscarini, "Decomposition of a hypergraph by partial-edge separators," *Theoretical Computer Science*, vol. 237, pp. 57–79, 2000.

[17] F. M. Malvestuto and E. Pourabbas, "Customized answers to summary queries via aggregate views," *Proceedings of the 16th SSDBM Conference*, pp. 193–202, 2004.

[18] R. Stone and A. Brown, *A Computable Model for Economic Growth, A Programme for Growth No.1.* London: Chapman and Hall, 1962.

[19] R. E. Tarjan and M. Yannakakis, "Simple linear-time algorithms to test chordality of graphs, test acyclicity of hypergraphs, and selectively reduce hypergraphs," *SIAM J. on Computing*, vol. 13, pp. 566–579, 1984.

APPENDIX

Section A

The *entropy* of a distribution $p(x)$ is the nonnegative functional

$$H[p] = -\sum_x p(x)\ log(p(x))$$

the summation being extended over all tuples x in the support of $p(x)$. It is well-known that $H[p]$ is always less than or equal to $log\ size(X)$. Given a distribution $\pi(x)$ whose support contains the support of $p(x)$, the *cross-entropy* (or "I-divergence"or "discrimination information" or "Kullback-Leibler distance") between $p(x)$ and $\pi(x)$ is the nonnegative functional

$$D[p,\pi] = \sum_x p(x)\ log\frac{p(x)}{\pi(x)}$$

the summation being extended over all tuples x in the support of $p(x)$. It is well-known that $D[p,\pi] = 0$ if and only if $p(x) = \pi(x)$. Let us assume that, for a subset Z of X, $q(z)$ is a distribution whose support contains the support of $p(z)$. Then, for $\pi(x) = \frac{q(z)}{size(X-Z)}$, we have

$$D[p,\pi] = logsize(X-Z) - H[p] - \sum_z p(z)\ logq(z).$$

Finally, suppose that $p(x)$ is an extension of a consistent set of distributions $p_1(x_1), \ldots, p_n(x_n)$. If Z is a (possibly empty) subset of X_i for some i, then

$$\sum_z p(z)\ logq(z) = \sum_z p_i(z)\ logq(z) = const$$

and, hence, minimizing $D[p,\pi]$ is the same as maximizing $H[p]$.

Section B

A *hypergraph* with vertex set X is a nonempty collection $\mathbf{H}$ of nonempty subsets of X, which are called *edges* of $\mathbf{H}$ [4] and whose union recovers X. A *partial edge* of $\mathbf{H}$ is a nonempty set of vertices that is contained in some edge of $\mathbf{H}$. A *cover* of $\mathbf{H}$ is a hypergraph $\mathbf{K}$ with vertex set X such that each edge of $\mathbf{H}$ is a partial edge of $\mathbf{K}$. Let W be a nonempty subset of X. The *subhypergraph* of $\mathbf{H}$ *induced*by W, denoted by $\mathbf{H}(W)$, is the hypergraph with vertex set W, whose edges are exactly the maximal (with respect to set-inclusion) intersections of W with the edges of $\mathbf{H}$. A *path* is a sequence of egdes such that every two consecutive edges have a nonempty intersection. Two vertices are *connected* if they belong respectively to the first edge and to the last edge of a path. The *connected components* of a hypergraph are its subhypergraphs induced by maximal sets of pairwise-connected vertices. A hypergraph is *connected* if it has exactly one connected component. Two connected vertices are *separated* by a set S of vertices if neither belongs to S and they belong to distinct connected components of $\mathbf{H}(X-S)$. A partial edge S is a *separator* if there exist two vertices that are separated by S but are not separated by any proper subset of S. Let $\mathbf{H}$ be a connected hypergraph. The *intersection graph* of $\mathbf{H}$ is the ordinary graph whose nodes correspond one-to-one to and are labelled by the edges of $\mathbf{H}$, and two distinct nodes of G are joined by an arc if their labels have a nonempty intersection. Moreover, if (u,v) is an arc of G and A and B are the labels of the nodes u and v, then the arc (u,v) is labelled by $A \cap B$. A spanning tree J of G is a *join tree* of $\mathbf{H}$ if, for every two nodes of J, the intersection of its labels is contained in the label of each node along the (unique) path in J that connects the two nodes. The hypergraph $\mathbf{H}$ is acyclic if there exists a join tree of $\mathbf{H}$ [4]. If this is the case, then for every join tree J of $\mathbf{H}$, the separators of $\mathbf{H}$ are exactly the labels of arcs of J. Several other equivalent definitions of acyclicity exist [4].