Relational algebra operations and sizes of relations

D. Gardy, C. Puech

INRIA

CENTRE DE ROCQUENCOURT

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
B.P. 105
78153 Le Chesnay Cedex
France
Tél. (3) 954 90 20

Rapports de Recherche

N° 317

RELATIONAL ALGEBRA OPERATIONS AND SIZES OF RELATIONS

Danièle GARDY
Claude PUECH

Juillet 1984

RELATIONAL ALGEBRA OPERATIONS AND SIZES OF RELATIONS (extended abstract)

Danièle GARDY and Claude PUECH

Laboratoire de Recherche en Informatique,
Bat 490, Université de Paris-Sud,
91405 Orsay Cédex,
France.

ABSTRACT

Abstract: *The performance of relational algebra operations is highly related to the sizes of the relations involved (sizes of initial relations as well as sizes of intermediate ones); in particular, any kind of knowledge on the sizes of intermediate relations is very useful with respect to query optimization as it can lead to order the basic operations involved in a particular query in a more efficient way. We present here a systematic way of estimating the sizes of the relations obtained from initial ones by applying relational algebra operations (projections, selections, intersections, unions, differences, equijoins, semijoins,...). Our approach relies on the description of relations by means of generating functions, and on the translation of the operations on relations into operators on the associated generating functions.*

Résumé: *Dans les systèmes de gestion de bases de données relationnelles, les performances des requêtes de l'algèbre relationnelle dépendent fortement des tailles des relations considérées. En particulier, la connaissance des tailles des relations intermédiaires permet d'ordonner les opérations composant une requête de manière à "optimiser" les temps de réponse. Ce travail présente une méthode systématique d'évaluation des tailles de relations obtenues par application d'une opération de l'algèbre (projection, sélection, intersection, union, jointure) à des relations de la base de données. Nous utilisons une fonction génératrice pour décrire les relations possibles de schéma donné et traduisons les opérations de l'algèbre relationnelle en opérateurs sur ces fonctions génératrices.*

1. INTRODUCTION.

As Relational Data Base Systems are now commercially available, the importance of the so-called *Query Optimization* is more and more evident. Some important work has already been devoted to the subject: [2, 3, 4, 10, 11, 13, 15, 16, 17, 18, 20, 21] , but most of it is based on heuristics and empirical observations.

As many authors pointed out, the size of the relations involved is one of the most important parameters; many queries can be decomposed in several ways into more elementary operations, and the overall performance of the query is highly dependent on the sizes of the intermediate relations produced by these intermediate operations; for example, as is well known, selection and projection operators should be applied as soon as possible as they provide a result whose cardinality is smaller than the one of their operands.

In this paper, we study the general problem of estimating the size of relations obtained as results of the basic relational algebra operators as a function of the size of their operands. Some related work can be found in [5, 6, 7, 8, 9, 16]

The originality of the proposed method (use of generating functions for describing relations) lies in its "generality": we propose to associate to every relation scheme and its possible dependencies † a generating function, and (as far as possible) to every relational algebra operator an operator acting on the generating functions associated on the operands; moreover, we try to give a systematic method for associating relations and generating functions ††.

This approach can be useful in two different ways:

(i) we could apply the machinery in an automatic way to produce the generating functions; then, deal with them by using some formal system in order to produce the distribution of the sizes of the results.

(ii) we can try to find classes of relations for which the generating functions have simple forms (usually they factorize) so as to be able to pull the calculations to the end and obtain closed formulae for the distribution of the sizes of results, its mean, variance ...

In the present paper, we develop point (ii). It is worth noting that although, for simplicity, we present the results either for "free" relations (no functional dependency) or for relations with a single functional dependency, some (but not all) more general cases can be solved as well with the same techniques.

The plan of the paper is the following: in Section 2, we give our probabilistic model(s) of relations, and show how to associate generating functions to relations; in Section 3, we deal with projections (the results obtained in this section

† see [12] or[19] for definitions of relational data base theory.
†† in the sequel, for the sake of simplicity, we use "relation" instead of "relation scheme and its possible dependencies"

can be useful in other domains as well: in *Data Analysis* projections of the initial multidimensional data on well chosen subspaces are used to obtain accurate "summaries" of the data; visually meaningful projections in two or three dimensions are also used in *Graphics* outputs); in Section 4, we deal with binary relational operators (intersection, union, difference) whose definition doesn't give any particular role to any particular attribute (as a consequence, the results are valid under any kind of hypothesis on the dependencies); in Section 5, we study equijoins and semijoins, whose importance in query optimization is well known; the last section gives directions for future research.

As regards the probabilistic hypotheses, let us mention that our hypotheses are, for finite domains, quite general; although the expressions obtained are more complicated in the "general" case (probability p_i for attribute X to have value i) the derivations are not very different from those for the uniform case; as was pointed out several times (see, for example[14]) skewed distributions of attribute values often arise in practice.

2. DESCRIPTION OF RELATIONS BY MEANS OF GENERATING FUNCTIONS.

We first give our probabilistic hypotheses on relations: in the sequel, we suppose that:

(i) distinct attributes of a record (tuple) are independent;

(ii) the probability of a relation is proportional to the probabilities of its records.

In other words, if R is a relation with $R[A_1,A_2,\cdots,A_n]$ as scheme, the probability of R is evaluated as:

$$p(R) = k \prod_{l \text{ tuple of } R} p(l) = k \prod_{l \text{ tuple of } R} \prod_{a_i \in l} p(a_i)$$

where the a_i are the components of tuple l, and k is a "normalization" constant which makes P a probability.

Let us mention that our method can be applied easily with a slightly different hypothesis (iii), instead of (ii):

(iii) the probability of relation R is evaluated as:

$$p(R) = k \prod_{l \text{ tuple of } R} p(l) \prod_{l \text{ not tuple of } R} (1-p(l))$$

$$= k' \prod_{l \text{ tuple of } R} \frac{p(l)}{(1-p(l))}$$

which could be more appropriate in some circumstances.

The description of relations by means of generating functions is made by using the following elementary lemmas: lemma 1 is a simple rewriting of the definition of a relation but, as it doesn't make any assumption on the relation, it

can be used under all circumstances by an "automatic analyzer"; lemma 2 describes *free* relations (i.e. without any functional dependency); lemma 3 describes relations with a single functional dependency and lemma 4 is useful for the study of binary operators on relations.

Lemma 1: *The formal polynomial*

$$P(\xi) = k \sum_{[R]} \prod_{t \in R} p(t)\xi_t \quad \text{where} \quad k = \frac{1}{\sum_{[R]} \prod_{t \in R} p(t)}$$

describes all relations with given scheme $[R]$, *in the following sense: the coefficient of* $\xi_{t_1}\xi_{t_2} \cdots \xi_{t_l}$ *in* P *is equal to the probability of the relation whose records are* $t_1, t_2, \cdots, t_l$.

Lemma 2: *The formal polynomial*

$$P(\xi) = k \prod_{t \in D} (1+p(t)\xi_t) \quad \text{where} \quad k = \frac{1}{\prod_{t \in D} (1+p(t))}$$

describes all the possible free relations on D, the domain of tuples.

Lemma 2': *The formal polynomial*

$$P(\xi,\eta) = k \prod_{t_X \in D_X} \prod_{t_Y \in D_Y} (1+p(t_X)p(t_Y)\xi_{t_X}\eta_{t_Y})$$

where $k = \dfrac{1}{\prod_{t_X \in D_X} \prod_{t_Y \in D_Y} (1+p(t_X)p(t_Y))}$ *describes all possible free relations* $R(X,Y)$ *on* $D=D_X \times D_Y$, *in the following sense: the coefficient of* $\xi_{t'_1}\eta_{t''_1}\xi_{t'_2}\eta_{t''_2} \cdots \xi_{t'_n}\eta_{t''_n}$ *in* $P(\xi,\eta)$ *is equal to the probability of the relation whose records are:* $t'_1t''_1, t'_2t''_2, \cdots, t'_nt''_n$.

Lemma 3: *The formal polynomial*

$$P(\xi,\eta) = k \prod_{t_X \in D_X} (1+p(t_X)\xi_{t_X} \sum_{t_Y \in D_Y} p(t_Y)\eta_{t_Y}) \quad \text{where} \quad k = \frac{1}{\prod_{t_X \in D_X} (1+p(t_X))}$$

describes all relations $R[X,Y]$ *with functional dependency* $X \rightarrow Y$.

Remark: Due to the functional dependency $X \rightarrow Y$ which is translated into the expression $\xi_{t_X} \sum p(t_Y)\eta_{t_Y}$, all the ξ_{t_X} are different in one monomial, which is not the case in Lemma 2'.

Lemma 4: *Let $P(\xi,\eta)$ be the formal polynomial associated to relation scheme $R[X,Y]$, and $Q(\xi',\zeta)$ the formal polynomial associated to relation scheme $S[X,Z]$; then, the formal polynomial:*

$$\Pi(\xi,\xi',\eta,\zeta) = P(\xi,\eta)Q(\xi',\zeta)$$

describes all couples of relations with these relation schemes.

Remark: In the lemmas X,Y,Z denote either an attribute or a set of attributes.

Notation: In the sequel we use the following notation:

$$\varphi(R_1{:}x_1,\cdots,R_i{:}x_i / S_1{:}y_1,\cdots,S_j{:}y_j) = \sum p_{r_1,\cdots,r_i;s_1,\cdots,s_j} x_1^{r_1}\cdots x_i^{r_i} y_1^{s_1}\cdots y_j^{s_j}$$

where the sum is taken over all $r_1,\cdots,r_i;s_1,\cdots,s_j$, and $p_{r_1,\cdots,r_i;s_1,\cdots,s_j}$ is the probability for relation R_1 to have size r_1, ... , for relation R_i to have size r_i, conditionned by: $|S_1| = s_1$, ... , $|S_j| = s_j$ ($|S_i|$ is the size of relation S_i).
As a consequence:

$$\varphi(R_1{:}x_1,\cdots,R_i{:}x_i / S_1{:}y_1,\cdots,S_j{:}y_j) = \frac{\varphi(R_1{:}x_1,\cdots,R_i{:}x_i,S_1{:}y_1,\cdots,S_j{:}y_j)}{\varphi(S_1{:}y_1,\cdots,S_j{:}y_j)}$$

3. PROJECTIONS.

We consider here relations $R[X,Y]$ with two (sets of) attributes X and Y, and study the size of the projection of R on the (set of) attribute(s) Y, denoted by $\Pi_Y(R)$.

The main result of this section is the following theorem:

Theorem 1: *The generating function of the sizes of R and its projection $\Pi_Y(R)$: $\varphi(R;x,\Pi_Y(R):y)$ is obtained from $P(\xi,\eta)$ (cf lemma 2' and lemma 3) by the following transform:*
- replace each ξ_{t_X} ($t_X \in D_X$) by x;
- then, for each $t_Y \in D_Y$, replace $\eta^{\alpha}_{t_Y}$ by 1 if $\alpha = 0$, y otherwise ($\alpha \geq 1$).

We give here a few consequences of theorem 1. As usual, means and variances are obtained by differentiation of the generating functions.

Corollary 1: *The probability, for a free relation $R[X,Y]$ of size l to have a projection $\Pi_Y(R)$ of size r is equal to* †:

$$\frac{[x^l\alpha^r]\prod_{i=1}^{d_X}[1-\alpha+\alpha\prod_{j=1}^{d_Y}(1+p_i\bar{p}_j x)]}{[x^l]\prod_{i=1}^{d_X}\prod_{j=1}^{d_Y}(1+p_i\bar{p}_j x)]}$$

† we denote by $[x^l]f(x)$ the coefficient of x^l in the Taylor expansion of f

(where the p_i and $\bar{p}_j$ are the respective probabilities of components t_X and t_Y of tuples t). This probability can also be expressed as:

$$\sum_{k=0}^{r}(-1)^{r-k}\binom{d_Y-k}{r-k}\frac{[x^l]\sum_{1\le i_1<\cdots<i_k\le d_X}\prod_{m=1}^{k}\prod_{j=1}^{d_Y}(1+p_{i_m}\bar{p}_j x)]}{[x^l]\prod_{i=1}^{d_X}\prod_{j=1}^{d_Y}(1+p_i\bar{p}_j x)]}$$

As for the uniform case (same probability for every attribute value),

Corollary 2: *Under the uniform hypothesis, the probability for a free relation $R[X,Y]$ of size l to have a projection $\Pi_Y(R)$ of size r is equal to:*

$$\frac{\binom{d_Y}{r}}{\binom{d_Xd_Y}{l}}\sum_{k=0}^{r}(-1)^{r-k}\binom{r}{k}\binom{kd_X}{l}$$

The mean of the distribution of the sizes of projections is equal to:

$$d_Y\left[1-\frac{\binom{d_Xd_Y-d_X}{l}}{\binom{d_Xd_Y}{l}}\right]$$

and its variance:

$$d_Y^2\left[\frac{\binom{d_Xd_Y-2d_X}{l}}{\binom{d_Xd_Y}{l}}-\frac{\binom{d_Xd_Y-d_X}{l}^2}{\binom{d_Xd_Y}{l}^2}\right]+d_Y\left[\frac{\binom{d_Xd_Y-d_X}{l}}{\binom{d_Xd_Y}{l}}-\frac{\binom{d_Xd_Y-2d_X}{l}}{\binom{d_Xd_Y}{l}}\right]$$

For relations with a single functional dependency, we can prove results such as:

Corollary 3: *The probability, for a relation $R[X,Y]$ with a single functional dependency $X\to Y$, of size l, to have a projection on Y of size r is given by the expression:*

$$\sum_{k=1}^{r}(-1)^{r-k}\binom{d_Y-k}{r-k}\sum_{1\le i_1<\cdots<i_k\le d_Y}(\bar{p}_{i_1}+\cdots+\bar{p}_{i_k})^l$$

Corollary 4: *Let $R[X,Y]$ be a relation with a single functional dependency, r its size. Under the uniform hypothesis, the distribution of the size of $\Pi_Y(R)$ has mean:*

$$d_Y\left(1-\left(1-\frac{1}{d_Y}\right)^r\right)$$

and variance:

$$d_Y^2((1-\frac{2}{d_Y})^r-(1-\frac{1}{d_Y})^{2r}) + d_Y((1-\frac{1}{d_Y})^r-(1-\frac{2}{d_Y})^r)$$

Corollary 5: *Let $R[X,Y]$ be a relation with a single functional dependency, r its size. Under the non-uniform hypothesis , the distribution of the size of $\Pi_Y(R)$ has mean:*

$$\sum_{t_Y \in D_Y} [1-(1-p(t_Y))^r]$$

and variance:

$$\sum_{t_Y \in D_Y} (1-p(t_Y))^r - [\sum_{t_Y \in D_Y} (1-p(t_Y))^r]^2 + \sum_{t_Y \in D_Y, t'_Y \in D_Y, t_Y \neq t'_Y} [1-p(t_Y)-p(t'_Y)]^r$$

4. INTERSECTIONS, UNIONS, DIFFERENCES, SELECTIONS.

We suppose in this section that R and S are two relations with the same relation scheme over a domain D of size d. Using lemma 4 to obtain a formal description of the couple (R,S), we can prove:

Theorem 2: *Let Λ be one of the following binary operators on relations: intersection ($\cap$), union ($\cup$), symetric difference (∇), complement (-). The generating function*

$$\varphi(R:x,S:y,\Lambda(R,S):z).$$

is obtained from the generating function $f(t_1,\ldots,t_d,t'_1,\ldots,t'_d)$ by substituting, for each i $(1 \leq i \leq d)$, a monomial $x^\alpha y^\beta z^\gamma$ to $t_i^k t'^l_i$, according to the rules given by Table 1.

	$R \cap S$	$R \cup S$	$R \nabla S$	$R-S$
$t_i^0 t'^0_i$	1	1	1	1
$t_i^1 t'^0_i$	xz	x	xz	xz
$t_i^0 t'^1_i$	yz	y	yz	y
$t_i^1 t'^1_i$	xyz	xyz	xy	xy

Table 1: *Substitutions.*

From this theorem several corollaries can be obtained for the distributions (and means and variances) of sizes of resulting relations.

If needed (for example, for the study of the composition of operations) a similar theorem could be obtained giving the formal polynomial associated to the relation $\Lambda(R,S)$. Here we need only the generating functions for the sizes of $\Lambda(R,S)$.

As for the case of selections, it can be reduced to an intersection problem.

5. EQUIJOINS AND SEMIJOINS.

Let $R[X,Y]$ and $S[X,Z]$ be two distinct relation schemes, $R \mid\times\mid S$ be the equijoin of R and S on attribute X, and $R \mid\times S$ the semijoin of R and S, i.e. the projection on the attributes X and Y of $R \mid\times\mid S$ (or the equijoin of R and $\Pi_X(S)$ on attribute X).

The basic transforms on generating functions associated to equijoins and semijoins are described in theorem 3.

Theorem 3: *The generating functions relative to equijoin size:*

$$\varphi(R:x,S:y,R \mid\times\mid S:z)$$

and semijoin size:

$$\varphi(R:x,S:y,R \mid\times S:z)$$

are obtained from the formal polynomial $\Pi(\xi,\xi',\eta,\zeta)$ associated to the couple $R\times S$ (cf Lemma 4) by the following transforms:

- *first, for each t_Y of D_Y, and each t_Z of D_Z, replace η_{t_Y} by x and ζ_{t_Z} by y;*
- *then, for each t_X of D_X, replace $(\xi_{t_X})^k(\xi'_{t_X})^l$:*
  - *by x^{kl}, in the case of equijoins,*
  - *by x^k if $l>0$, 1 otherwise, in the case of semijoins.*

Under the uniform hypothesis, the generating function for semijoin size can be expressed as a sum according to the following corollary:

Corollary 6: *Under the uniform hypothesis, the generating function for the size of the semijoin $R \mid\times S$ of $R[X,Y]$ and $S[X,Z]$ is given by:*

$$\varphi(R:x,S:y,R \mid\times S:z) = \sum_{0\le m\le l\le d_X} (-1)^{m+l} \binom{d_X}{l} \binom{l}{m} f_l(x,y) g_m(z)$$

where $f_l(x,y)$ is obtained from the formal polynomial $f_R((x_i),(y_j))$ describing relation R by substituting y to each y_j, x to l of the x_i, 1 to the other x_i; $g_m(z)$ is obtained from the formal polynomial $g_S((x'_i),(z_k))$ describing relation S by substituting z to each z_k, 1 to m of the x'_i, 0 to the other x'_i.

Corollary 6 makes no assumption on the relation schemes for R and S. The only hypothesis is the uniform distribution of attribute values in their domains which guarantees that f_R and g_S are symetric functions of the x_i and x'_i respectively.

If we suppose, moreover, that R and S are either free relations or relations with a single functional dependency, we can obtain explicit formulae for the generating function of the semijoin size.

Corollary 7: *Let $R[X,Y]$ and $S[X,Z]$ be either free relations or relations with a single functional dependency. Under the uniform hypothesis, the generating function*

$$\varphi(R:x,S:y,R\,|\times S:z)$$

for the semi-join is given by table 2 (where $X\dagger Y$ means there is no functional dependency between X and Y).

R	S	φ
$X\dagger Y$	$X\dagger Z$	$[(1+y)^{d_Y}+(1+xy)^{d_Y}[(1+z)^{d_Z}-1]]^{d_X}$
$X\dagger Y$	$X\to Z$	$[(1+y)^{d_Y}+d_Z z(1+xy)^{d_Y}]^{d_X}$
$X\dagger Y$	$Z\to X$	$\sum_{k=0}^{d_X}\binom{d_X}{k}(1+kz)^{d_Z}(1+xy)^{kd_Y}[(1+y)^{d_Y}-(1+xy)^{d_Y}]^{d_X}$
$X\to Y$	$X\dagger Z$	$[1+d_Y y+(1+d_Y xy)[(1+z)^{d_Z}-1]]^{d_X}$
$X\to Y$	$X\to Z$	$[1+d_Y y+d_Z z(1+d_Y xy)]^{d_X}$
$X\to Y$	$Z\to X$	$\sum_{k=0}^{d_X}\binom{d_X}{k}(1+kz)^{d_Z}(1+d_Y xy)^k[d_Y y(1-x)]^{d_X-k}$
$Y\to X$	$X\dagger Z$	$\sum_{k=0}^{d_X}\binom{d_X}{k}[(1+z)^{d_Z}-1]^k(1+d_X y+ky(x-1)]^{d_Y}$
$Y\to X$	$X\to Z$	$\sum_{k=0}^{d_X}\binom{d_X}{k}d_Z z^k[1+d_X y+ky(x-1)]^{d_Y}$
$Y\to X$	$Z\to X$	$\sum_{0\le k\le l\le d_X}(-1)^{k+l}\binom{d_X}{l}\binom{l}{k}(1+kz)^{d_Z}[1+d_X y+ly(x-1)]^{d_Y}$

Table 2: *Generating function for semijoin size.*

For the mean and variance of the distribution, we can prove:

Corollary 8: *Let $R[X,Y]$ and $S[X,Z]$ be either free relations or relations with a single functional dependency. Let r denote the size of R, s the size of S, and:*

$$\alpha=\frac{\binom{d_X d_Z-d_Z}{s}}{\binom{d_X d_Z}{s}}\qquad \alpha_1=\left(1-\frac{1}{d_X}\right)^s$$

$$\beta = \frac{\binom{d_X d_Z - 2d_Z}{s}}{\binom{d_X d_Z}{s}} \qquad \beta_1 = \left(1 - \frac{2}{d_X}\right)^s$$

Under the uniform hypothesis, the mean value of the size of the semijoin $R \ltimes S$ (which does not depend on the relation scheme for S) is given by Table 3, and the variance of the distribution of semijoin sizes is given by Table 4.

$X^{\dagger}Y$	$X \rightarrow Y$	$Y \rightarrow X$
$r(1-\alpha)$	$\frac{rs}{d_X}$	$r(1-\alpha_1)$

Table 3: *Mean size of semijoin.*

R	S	*Variance*
$X^{\dagger}Y$	$X^{\dagger}Z$	$r\alpha\frac{(dx-1)d_Y+r(d_Y-1)}{d_X d_Y-1} - r^2\alpha^2+r(r-1)\frac{d_X(d_Y-1)}{d_X d_Y-1}\beta$
$X^{\dagger}Y$	$X \rightarrow Z$	$\frac{rs}{d_X}(1-\frac{s}{d_X})\frac{d_X d_Y-r}{d_X d_Y-1}$
$X^{\dagger}Y$	$Z \rightarrow X$	$r\alpha_1\frac{(d_X-1)d_Y+r(d_Y-1)}{d_X d_Y-1} - r^2\alpha_1^2+r(r-1)\frac{d_X(d_Y-1)}{d_X d_Y-1}\beta_1$
$X \rightarrow Y$	$X^{\dagger}Z$	$r\alpha+r(r-1)\beta-r^2\alpha^2$
$X \rightarrow Y$	$X \rightarrow Z$	$\frac{rs}{d_X-1}(1-\frac{r}{d_X})(1-\frac{s}{d_X})$
$X \rightarrow Y$	$Z \rightarrow X$	$r\alpha_1+r(r-1)\beta_1-r^2\alpha_1^2$
$Y \rightarrow X$	$X^{\dagger}Z$	$r(1+\frac{r-1}{d_X})\alpha-r^2\alpha^2+r(r-1)(1-\frac{1}{d_X})\beta$
$Y \rightarrow X$	$X \rightarrow Z$	$\frac{rs}{d_X}(1-\frac{1}{d_X})$
$Y \rightarrow X$	$Z \rightarrow X$	$r(1+\frac{r-1}{d_X})\alpha_1-r^2\alpha_1^2+r(r-1)(1-\frac{1}{d_X})\beta_1$

Table 4: *Variance of the distribution of semijoin size.*

Note that when the domains are large the variance is approximately equal to $\frac{rs}{d_X}$, except in the case $X^{\dagger}Y$, $X^{\dagger}Z$ for which the approximate value is: $\frac{rs}{d_X d_Y}$

In the uniform case, the distribution of equijoin sizes, its mean and variance, can also be expressed by closed formulae.

Corollary 9: *Let $R[X,Y]$ and $S[X,Z]$ be either free relations or relations with a single functional dependency. Under the uniform hypothesis, the generating function for the size of the equijoin:*

$$\varphi(R{:}x,S{:}y,R \mid\times\mid S{:}z)$$

is given by Table 5 (which has to be completed by symmetry).

R	S	φ
$X\uparrow Z$	$X\uparrow Y$	$[\sum_{k=0}^{d_Y}\binom{d_Y}{k}y^k(1+x^kz)^{d_Z}]^{d_X}$
$X\to Z$	$X\uparrow Y$	$[(1+y)^{d_Y}+d_Z z(1+xy)^{d_Y}]^{d_X}$
$X\to Z$	$X\to Y$	$(1+d_Y y+d_Z z+d_Y d_Z xyz)^{d_X}$
$Z\to X$	$X\uparrow Y$	$\sum_{k=0}^{d_X d_Y} y^k \sum_{0\le u_i\le d_Y; u_1+\cdots+u_{d_X}=k}\binom{d_Y}{u_1}\cdots\binom{d_Y}{u_{d_X}}[1+z\sum_{i=1}^{d_X}x^{u_i}]^{d_Z}$
$Z\to X$	$X\to Y$	$\sum_{k=0}^{d_X}\binom{d_X}{k}(d_Y y)^k[1+d_X z+kz(x-1)]^{d_Z}$
$Z\to X$	$Y\to X$	$\sum_{k=0}^{d_Y}\binom{d_Y}{k}\sum_{0\le u_i; u_1+\cdots+u_{d_X}=k}\binom{k}{u_1\ \cdots\ u_{d_X}}[1+z\sum_{i=1}^{d_X}x^{u_i}]^{d_Z}$

Table 5: *Generating function for equijoin size.*

Corollary 10: *Let $R[X,Y]$ and $S[X,Z]$ be either free relations or relations with a single functional dependency. Under the uniform hypothesis, the distribution of equijoin size has mean:*

$$\frac{rs}{d_X}$$

Its variance is given in Table 6 (which is symmetric).

Although the results become much more intricate, similar results can be obtained in the non uniform case. For example:

Corollary 11: *Let $R[X,Y]$ and $S[X,Z]$ be two relations with functional dependencies $X\to Y$ and $X\to Z$.*
Under the non-uniform assumption:

$$\varphi[R{:}x,S{:}y,R \mid\times\mid S{:}z] = \prod_{t_X\in D_X}(1+p(t_X)(y+z)+p(t_X)^2xyz)$$

R	S	*Variance*
$X^{\dagger}Z$	$X^{\dagger}Y$	$\frac{rs}{d_X} \times \frac{d_X-1}{d_X} \times \frac{d_X d_Y - r}{d_X d_Y - 1} \times \frac{d_X d_Z - s}{d_X d_Z - 1}$
$X^{\dagger}Z$	$X \rightarrow Y$	$\frac{rs}{d_X} \times \frac{d_X d_Z - s}{d_X d_Z - 1} \times \frac{d_X - r}{d_X}$
$X^{\dagger}Z$	$Y \rightarrow X$	$\frac{rs}{d_X} \times \frac{d_X-1}{d_X} \times \frac{d_X d_Z - s}{d_X d_Z - 1}$
$X \rightarrow Z$	$X^{\dagger}Y$	$\frac{rs}{d_X} \times \frac{d_X d_Y - r}{d_X d_Y - 1} \times \frac{d_X - s}{d}$
$X \rightarrow Z$	$X \rightarrow Y$	$\frac{rs}{d_X - 1} \times \frac{d_X - r}{d_X} \times \frac{d_X - s}{d_X}$
$X \rightarrow Z$	$Y \rightarrow X$	$\frac{rs}{d_X} \times \frac{d_X - s}{d_X}$
$Z \rightarrow X$	$X^{\dagger}Y$	$\frac{rs}{d_X} \times \frac{d_X-1}{d_X} \times \frac{d_X d_Y - r}{d_X d_Y - 1}$
$Z \rightarrow X$	$X \rightarrow Y$	$\frac{rs}{d_X} \times \frac{d_X - r}{d_X}$
$Z \rightarrow X$	$Y \rightarrow X$	$\frac{rs}{d_X} \times \frac{d_X-1}{d_X}$

Table 6: *Variance for equijoin size.*

6. DIRECTIONS FOR FUTURE RESEARCH.

We have now tools for obtaining the distribution of the sizes of results of relational algebra operations in several well understood cases. We intend to pursue our work in the following directions:

i) evaluate, using our results, the overall performance of basic relational queries, taking into account several possible implementations of the relations (hash tables ...) and the "physical" cost of retrieval.

ii) study the effect of "cascades" of operations on the sizes of relations;

iii) study the influence of the size of the domain (the approach can also be used to study infinite domains);

iv) try to deal with more complex dependencies (several dependencies, multivalued dependencies ...);

iv) test the usefulness of our approach for implementing an automatic "size_of_result" analyzer, a first step towards a "query optimizer"...

Let us also mention that some of our results seem useful for analyzing the size of compacted files obtained, for example, by the method described in [1].

References

1. F. Bancilhon, Ph. Richard, and M. Scholl, *On line processing of compacted relations*, Proc. VLDB 82, Mexico 1982.
2. M.W. Blasgen and K.P. Eswaran, "On the Evaluation of Queries in Relational Data Base Systems," RJ1745 (April 1976). IBM Research Report, IBM Research Center, San Jose
3. S. Christodoulakis, *Estimating Block Transfers and Join Sizes*, Proc. SIGMOD 83, San Jose, Cal. 1983.
4. W.W. Chu and P. Hurley, "Optimal query processing for distributed database systems," *IEEE Transactions on Computers* C-31(9)(1982).
5. R. Demolombe, *Estimation of the number of tuples satisfying a query expressed in Predicate Calculus language* , Proc. VLDB 80 1980.
6. R. Demolombe, "How to improve performance of relational DBMS," pp. 229-233 in *Proc. IFIP 83*, ed. Mason R.E.A.,Elsevier Science Publishers (1983).
7. D. Gardy, *Evaluation de résultats d'opérations de l'algèbre relationnelle*, Thèse de Troisième Cycle, Université de Paris-Sud, Orsay 1983.
8. E. Gelenbe and D. Gardy, "On the sizes of projections, I," *Information Processing Letters* 14(1)(1982).
9. E. Gelenbe and D. Gardy, *The size of Projections of Relations Satisfying a Functional Dependency*, Proc. VLDB 82, Mexico 1982.
10. L.R. Gotlieb, *Computing joins of relations*, Proc. SIGMOD 78 1978.
11. A.R. Hevner and S.B. Yao , "Query processing in distributed data base systems," *IEEE Transactions On Software Engineering*, (1979).
12. D. Maier, *The Theory of Relational Databases*, Computer Science Press (1983).
13. T.H. Merrett and Ekow Otoo, *Distributions models of relations*, Proc. VLDB 79, Rio de Janeiro 1979.
14. A.Y. Montgomery, Y.J. D'Souza, and S.B. Lee , "The cost of relational algebraic operations on skewed data: estimates and experiments," pp. 235-241 in *Proc. IFIP 83*, ed. Mason R.E.A.,Elsevier Science Publishers (1983).
15. Ph. Richard, *Evaluation of the size of a query expressed in relational algebra*, Proc. SIGMOD 81 1981.
16. A.S. Rosenthal, "Note on the expected size of a join," *SIGMOD Record* 11(4) pp. 19-25 (1981).
17. P. Griffiths Selinger, M.M. Astrahan , M.M. Chamberlin , R.A. Lorie, and T.G. Price, *Access path selection in a relational Database System*, Proc. ACM SIGMOD 1979.
18. J.M. Smith and P.Y.T. Chang, "Optimizing the performance of a relational algebra Database interface," *CACM* 18(10)(1975).

19. J.D. Ullmann, ***Principles of data base systems***, Computer Science Press (1980).

20. S.B. Yao, "An attribute based model for data base access cost analysis," *ACM TODS*, (1977).

21. S.B. Yao, "Optimization of Query Evaluation Algorithms," *ACM TODS* 4(2) pp. 133-155 (1979).

Imprimé en France

par

l'Institut National de Recherche en Informatique et en Automatique