Theoretical Computer Science 131 (1994) 375–414
Elsevier

Join sizes, urn models and normal limiting distributions

Danièle Gardy*

LRI, Université de Paris-Sud, CNRS U.A. 410, Bât. 490, 91405 Orsay Cedex, France

Communicated by M. Nivat
Received June 1991
Revised April 1993

Abstract

Gardy, D., Join sizes, urn models and normal limiting distributions, Theoretical Computer Science 131 (1994) 375–414.

We study some parameters of relational databases (sizes of relations obtained by a join) that can be described by generating functions on three variables, of the kind $\varphi(x, y, z)^d$. We modelize these parameters by suitable urn models and give conditions under which they asymptotically follow a gaussian distribution.

1. Introduction

A crucial problem in relational databases is that of query optimization, i.e. the choice of a strategy to compute the data required by an user. There are usually many ways to answer a query; a specialized module of the database system (the query optimizer) chooses among them according to the underlying data structures and system, and its global goal is generally to minimize some cost function [10]. The precise definition of the cost of a query depends on the database system; most of the optimizers use some common parameters, such as the number of disk transfers, the memory used, the amount of data transferred from site to site in a distributed system. The size of the data, either present in the database or computed during the evaluation of a query, is one of these parameters [2, 14]. We have chosen to study the sizes of the relations (i.e. structures used to store the data) obtained by the operations of the

Correspondence to: D. Gardy, Laboratoire PRiSM, Université de Versailles St-Quentin, 45 avenue des Etats-Unis, F-78035 Versailles Cedex, France. Email: Daniele.Gardy@prism.uvsq.fr.

* This work was partially supported by the PRC Mathématique-Informatique and by ESPRIT-II Basic Research Action Contract No. 3075 (project ALCOM).

SSDI 0304-3975(93)E0104-C

relational algebra, which is a widely used high-level language for relational databases. We presented a model for the evaluation of these sizes under some general conditions in collaboration with Puech [7, 8], and asymptotic results on projection sizes and some join sizes in [6]. This paper is a sequel to [6], and extends its results to several other cases of joins.

Our approach for studying the effect of a relational operation on the size of a relation can be summarized as follows. We associate with each relation a (total) generating function and with each operation on relations an operator on these functions, which describes its effect on the sizes of relations; we obtain a multivariate generating function marking the sizes of the initial relation(s) and of the resulting relation, and we use this function to study the distribution of the size of the derived relation when the sizes of the initial relations are known. This has led us to study in [6] bivariate generating functions which have the general form $\varphi(x, y)^d$, and the limiting distributions they define for large d. This paper is an extension of our former results to generating functions of three variables $\Phi(x, y, z) = \varphi(x, y, z)^d$, when there is no easy reduction to a problem in two variables. As a consequence, the technics we use become rather involved, and the length of the proofs justifies the presentation in two separate papers.

The present paper is organized as follows. In Section 2 we recall the definition of relations and of the operators of the relational algebra, and the way we associate generating functions with them. We also give there an interpretation of the sizes of the projections and joins in terms of urn models and occupancy problems. Section 3 presents our results. We give general conditions under which the join size asymptotically follows a normal limiting distribution. In terms of generating functions, this means that the probability distribution defined by the probability generating function $f(x) = [y^r z^s]\Phi(x, y, z)/[y^r z^s]\Phi(1, y, z)$, for $\Phi(x, y, z) = \varphi(x, y, z)^d$, tends towards a normal distribution for $d \to +\infty$, $r \approx Ad$ and $s \approx Bd$, and when the function φ has real positive coefficients and belongs to some general class. The functions used for joins are of the kind $\varphi(x, y, z) = \lambda_1(y) + \lambda_1(xy)(\lambda_2(z) - 1)$ for suitable functions λ_1 and λ_2 and of the kind $\varphi(x, y, z) = \sum_{k,l \geq 0} \alpha_{k,l} x^{kl} y^k z^l$. We then discuss what we have obtained and give indications for possible extensions in Section 4. Finally, we give the proofs of our asymptotic theorems in Section 5.

2. Urn models, semijoins and equijoins

2.1. Relational databases and operations

We present below some definitions and terminology relative to relational databases; we refer the reader to [13, 16] for a more complete presentation.

2.1.1. Relations

In a relational database, data are stored in tabular structures called *relations*. An instance of a relation, or more briefly a relation, is a set of points (*tuples*) in

a multidimensional space. The coordinates of the tuples are the *attributes* of the relation; each attribute takes its values in a finite domain. We shall use the following notation: $R[X, Y]$ is a relation R which has for attributes X and Y; attribute X has for domain D_X, and d_X is the size of D_X, i.e. the number of distinct values that attribute X may take; similar conventions hold for attribute Y.

2.1.2. Constraints on relations

Relations usually satisfy constraints, which have their origin in the data to be modelled. The simpler constraints are enforced simply by a suitable choice of the domains of attributes. Other constraints restrict the sets of tuples that can be formed. We shall not study all possible cases, but restrict ourselves to relations either without constraint or with a key. When there is no constraint at all between the tuples of a set, the relation is said to be *free*, and there is total independence between the values taken by the different tuples. We may also consider relations where an attribute is a *key*: in a given instance of the relation, the value of a tuple on this attribute uniquely determines its value on the other attributes, for all the tuples of the instance.

2.1.3. Relational algebra

The classical operators defined on relations are the set operations (union, intersection, difference) on two relations with the same (or compatible) attributes, the cartesian product of two relations, the selection of the tuples of a relation satisfying some condition, the projection of a relation on an attribute or on a set of attributes, and the equijoin of two relations.[1] For simplicity, and without loss of generality, we shall restrict ourselves to the case where the two relations involved in an equijoin have a unique common attribute.

The *projection* of a relation on a set of attributes is obtained by suppressing in each tuple the values on the attributes which do not belong to this set, then removing the duplicate tuples in the resulting relation. We give in Fig. 1 an instance of a relation $R[X, Y]$ and of its projection (noted $\pi_X(R)$) on attribute X.

R

X	Y
x_0	y_0
x_0	y_1
x_1	y_2

$\pi_X(R)$

X
x_0
x_1

Fig. 1. Projection of relation $R[X, Y]$ on attribute X.

[1] The relational algebra is redundant. For example, the equijoin can be defined using the cartesian product, the selection and the projection.

The *equijoin* of two relations $R[X, Y]$ and $S[X, U]$ on their common attribute X has three attributes X, Y and U; it is composed of all triples (x, y, u) such that (x, y) belongs to R and (x, u) belongs to S. This definition is easily extended to relations with more than two attributes. Up to a reordering of the attributes, the equijoin is a symmetrical operation: the equijoin of relations R and S is equal to the equijoin of relations S and R. Figure 2 presents instances of two relations $R[X, Y]$ and $S[X, U]$ and of their equijoin (denoted $R \bowtie S$) on attribute X.

The *semijoin* of two relations $R[X, Y]$ and $S[X, U]$ on their common attribute X (we use the notation $R \triangleright S$) does not belong to the relational algebra *stricto sensu*, but is useful enough to be often included in it. It is computed by discarding from relation R those tuples whose value on X does not appear in the X-column of relation S: $R \triangleright S = \{(x, y) \in R / \exists u\colon (x, u) S\}$. The semijoin is thus the composition of a projection and an equijoin: $R[X, Y] \triangleright S[X, U] = \pi_{XY}(R \bowtie S) = R \bowtie \pi_X(S)$. This operation is not symmetrical: the semijoin of R and S is *not* equal to the semijoin of S and R (see Fig. 3 for an example; the instances of relations R and S are as in Fig. 2).

We assume that the relations we consider have two (sets of) attributes. We shall work with a relation $R[X, Y]$ and a relation $S[X, U]$. Throughout the paper, X denotes the join attribute; when working on a semijoin we shall consider the semijoin $R \triangleright S$ of R and S, unless indicated otherwise.

2.2. *Urn models for relational operations*

The classical *occupancy problem* for urn models is defined as follows [11]: Given d distinguishable urns, we independently throw n balls into these urns. We are interested in the number of empty urns or, equivalently, in the number of urns with at

R

X	Y
x_0	y_0
x_0	y_1
x_1	y_2

S

X	U
x_0	z_0
x_0	z_1
x_1	z_1
x_2	z_2

$R \bowtie S$

X	Y	U
x_0	y_0	z_0
x_0	y_0	z_1
x_0	y_1	z_0
x_0	y_1	z_1
x_1	y_2	z_1

Fig. 2. Equijoin of the relations $R[X, Y]$ and $S[X, U]$ on attribute X.

$R \triangleright S$

X	Y
x_0	y_0
x_0	y_1
x_1	y_2

$S \triangleright R$

X	U
x_0	z_0
x_0	z_1
x_1	z_1

Fig. 3. Semijoins of R and S, and of S and R.

least one ball. In the most common case, the balls are equivalent, and the capacity of any urn is not bounded. Let $N_{n,k}$ be the number of ways of allocating n balls into exactly k urns among d (there are $d-k$ empty urns). Classical ideas of combinatorial enumeration (see e.g. [9]) provide an easy way to obtain the generating function $\Phi(x, y)$, where x marks the number of urns with at least one ball, and y the number of balls. (The function Φ can also be obtained by elementary enumeration methods, see [11] for such an approach.) The balls being equivalent, the "natural" choice is for a function Φ exponential in y: the order in which the n balls were thrown into the urns has no influence on the number of empty urns. However, the urns themselves are distinguishable and this suggests that Φ should be an ordinary generating function in x: We thus define $\Phi(x, y)=\sum_{n,k} N_{n,k} x^k y^n/n!$.

Let $\varphi(x, y)$ be the generating function describing what happens in any one urn, with x "marking" the event that there is at least one ball in this urn, and y counting the number of balls. Let v_i be the number of ways of choosing i balls. The order in which the balls are thrown into the urn does not matter, and we have that $\varphi(x, y)=1+x\sum_{i\geq 1} v_i y^i/i!$. However, the i balls are indistinguishable, hence there is just one way to choose a given number i of balls: $v_i=1$. This shows that

$$\varphi(x, y)=1+x(e^y-1).$$

The independence of the urns means that the function $\Phi(x, y)$ is the product of the d elementary generating functions $\varphi(x, y)$ associated with the urns, which gives

$$\Phi(x, y)=(1+x(e^y-1))^d.$$

Let us now compute the probability $p_{n,k}$ that n balls fall into exactly k urns among the d possible urns. The number of ways to allocate n balls into d urns, some of which may be empty, is d^n, and the probability that, after throwing n balls independently, exactly k urns among d contain at least one ball is equal to $N_{n,k}/d^n$. This probability can be recovered from the function Φ:

$$p_{n,k}=\left[x^k \frac{y^n}{n!}\right]\Phi(x, y) \Big/ \left[\frac{y^n}{n!}\right]\Phi(1, y).$$

If now the allocation of the balls into an urn is restricted in some way (urns of bounded capacity, distinguishable balls with probabilities, etc.), we use a function λ to sum up this information. The generating function $\varphi(x, y)$ marking the fact that the urn is not empty becomes $\varphi(x, y)=1+x(\lambda(y)-1)$. The global generating function $\Phi(x, y)$, where the variable x marks the number of urns which contain at least one ball and the variable y marks the total number of balls, is then

$$\Phi(x, y)=(1+x\,(\lambda(y)-1))^d. \tag{1}$$

In terms of relations, a generating function of the kind (1) corresponds to the projection of a relation $R[X, Y]$ on the attribute X, and the function $\lambda(t)$ is associated

with the underlying scheme of the relation (set of constraints that each instance must satisfy); see [5, 6] and Section 2.4 for a discussion of the relationship between the function λ and the relation scheme. The variable y marks the size of the initial relation, and the variable x marks its projection on attribute X.

More formally, we propose to associate an occupancy model with the projection of a relation $R[X, Y]$ on the attribute X as follows. We choose as many urns (d_X) as there are possible values for attribute X, and label each urn by a distinct value of the domain of X. If an instance of the relation R has n tuples, we throw n balls into the urns, according to a specified set of rules that differs according to the existence or absence of a key in relation R (see [6, pp. 225–227] for a justification in terms of generating functions):

If the relation R has for key the projection attribute X, an urn may contain at most one ball; if the attribute Y suppressed by the projection is key, we consider that the urns are of unbounded capacity, and that the balls are undistinguishable; finally, if the relation R is free, the balls are distinguishable, and the urns have a capacity of d_Y balls.

Each ball represents a tuple (x_0, y_0) and goes into the urn labelled by x_0. At the end of n trials, the number of urns with at least one ball is equal to the number of distinct X-values in the instance of the relation R, i.e. to the size of the projection of R on attribute X.

The semijoin and equijoin sizes can likewise be expressed in the general framework of urn models, and the generating functions of the parameters of interest can easily be computed. We present below two models that can be used to describe the equijoin and the semijoin.

We assume that we have two kinds of balls, say blue and red. The balls of a given color are thrown into d urns independently of each other and of the balls of the other color. After throwing specified numbers of red and blue balls, we assign a certain number of balls of a third color, say green, to the urns according to one of the two sets of rules below. The first extension of the classical model is defined as follows (see Fig. 4).

Model A:

- We throw into the urns a given number r of red balls and a given number s of blue balls.
- For each urn containing at least one blue ball, put as many green balls as there are red balls. The urns without balls or with balls of only one color do not receive any green ball.
- Count the total number of green balls.

The second extension is as follows (see Fig. 5).

Model B:

- We throw into the urns a given number r of red balls and a given number s of blue balls.
- For each urn where there are i red balls and j blue balls, put ij green balls in the urn. If an urn contains no balls, or balls of only one color, we put no green ball into this urn.
- Count the total number of green balls.

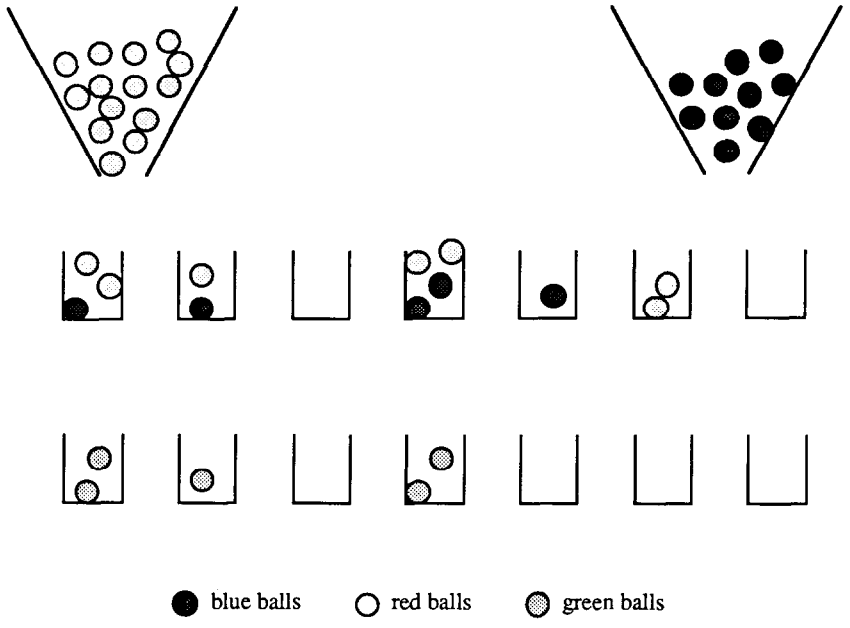

Fig. 4. Model A: semijoin.

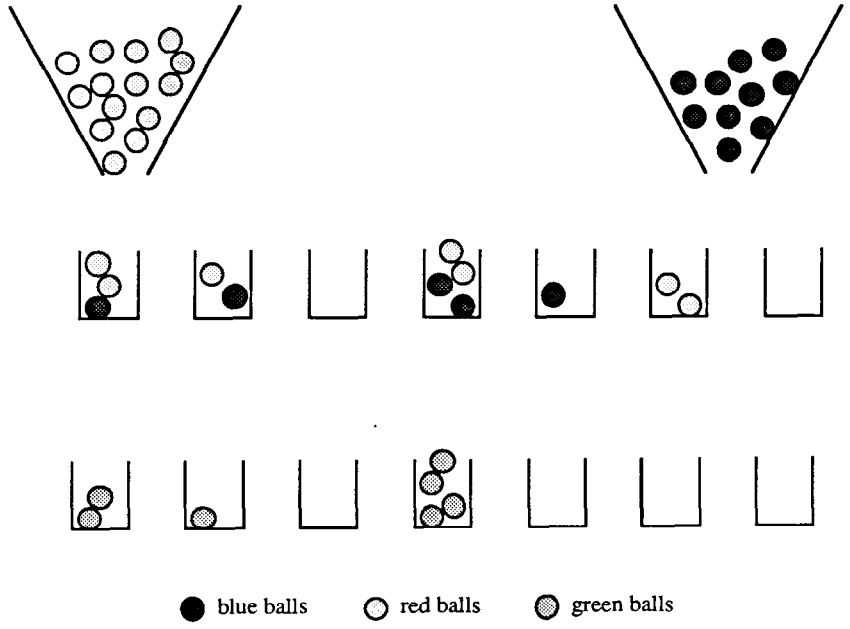

Fig. 5. Model B: equijoin.

These two urn models are easily associated with operations on relations as follows: Each type of balls corresponds to a relation, for example, red balls are associated with the relation R and blue balls with the relation S; a trial in which we throw r red balls corresponds to the choice of an instance of size r for the relation R. The exact rules for throwing red and blue balls (capacity of urns, distinguishable or indistinguishable balls) depend on the relation schemes and were given formerly for the projection. The number of urns is equal to the size d_X of the join attribute, and each urn is labelled by a distinct value of the domain D_X. The total number of green balls computed in model A is the size of the semijoin of R with S (recall that the semijoin is not a symmetrical operation). The number of green balls computed in model B is the size of the equijoin of R and S.

For the two urn models described above, and assuming undistinguishable balls and urns of unbounded size, the generating functions marking the initial numbers r and s of red and blue balls by the variables y and z, and the final number of green balls by the variable x, have a common form: they are equal to the dth power of a function with positive coefficients. We shall see in Section 2.4 that it is possible to choose the generating function (ordinary or exponential, counting or probability generating function, etc.) marking the size of the equijoin or semijoin of two relations R and S together with the sizes of these initial relations, in such a way that it also follows this common form.

As was the case for the projection, the way we throw red balls into any given urn can be summed up in a generating function $\lambda_R(t)=\sum_k a_k t^k$. A similar function $\lambda_S(t)=\sum_l b_l t^l$ describes the behavior of blue balls in any given urn. The generating function corresponding to model A (and to the semijoin size) is

$$\Phi(x, y, z)=(\lambda_R(y)+\lambda_R(xy)(\lambda_S(z)-1))^d.$$

To prove it, we first consider how red and blue balls can fall into a given urn: this is described by the product $\lambda_R(y)\lambda_S(z)$ (recall that y marks the red balls and z the blue balls). Either there is no blue ball, which gives the term $\lambda_R(y)$, or there is at least one blue ball, which corresponds to $\lambda_R(y)\,(\lambda_S(z)-1)$, and putting as many green balls into the urn as there are red balls gives $\lambda_R(xy)\,(\lambda_S(z)-1)$. The generating function associated with an urn is thus $\lambda_R(y)+\lambda_R\,(xy)\,(\lambda_S(z)-1)$, and the expression for Φ follows from the independence of the d urns.

The generating function associated with model B (equijoin size) is computed in a similar way. Once again, we first look at the way red balls and blue balls can fall into any given urn: this is described by $\lambda_R(y)\,\lambda_S(z)=\sum_{k,l\geq 0} a_k b_l y^k z^l$. The way green balls are then associated with this urn translates into the term $\sum_{k,l\geq 0} a_k b_l x^{kl} y^k z^l$, and the final form of Φ follows immediately:

$$\Phi(x, y, z)=\left(\sum_{k,l\geq 0} a_k b_l x^{kl} y^k z^l\right)^d.$$

2.3. Models and hypotheses

We present here the assumptions under which we shall estimate the size of the join of two relations of known sizes. The reader may find a more complete presentation in [6] and a justification in terms of relational database theory in [5, 8].

2.3.1. Independence assumptions

We assume that the relations satisfy some independence assumptions:

- *The coordinates of a tuple are independent*;
- *The tuples of a given relation are independent as far as this is compatible with the constraints on the relation (free relation or relation with a key)*;
- *Two relations $R[X, Y]$ and $S[X, U]$ are independent, unless otherwise indicated (in Section* 3.4).

The first condition simply states that the probability distribution on the cartesian product $D_X \times D_Y$ is the product of the probability distributions on domains D_X and D_Y. The second assumption means that the probability distribution on a relation R is proportional to the probability of each of its tuples: $\text{Prob}(R) = k \prod_{t \in R} \text{Prob}(t)$, for a constant k independent of R, and chosen according to the underlying constraints on the relation. The last condition just states that the probability distribution of a couple (R, S) is the product of the probabilities of R and S.

We should point out that, although the semijoin is the composition of a projection and an equijoin, the knowledge of the probability distributions of the sizes of the projection and the equijoin is not always sufficient to estimate the probability distribution of the semijoin: assumptions on the initial relations R and S (independence of tuples, distributions on attribute domains) may not be satisfied by either the projection $\Pi_X(R)$ or the equijoin $R \bowtie S$.

2.3.2. Probability distributions on attribute domains

All possible values of an attribute do not always appear with the same frequency in a relation. We assume that we know (or can compute, at least approximately) the probability distribution on each attribute domain.

We consider two kinds of distributions on a domain D of d elements according to whether the distribution is not too far from the uniform distribution or is strongly biased. Let $p_{i,d}$ be the probability of the ith element of D and assume that $d \to +\infty$. We obtain a sequence of distributions indexed by d; the two classes (Z) (for Zipf distribution) and (G) (for geometric distribution) are defined as follows:

- (Z) $\sum_{i=1}^{d} p_{i,d}^2 \to 0$ for $d \to +\infty$.
- (G) For each fixed i, $p_{i,d} \to p_i$ for $d \to +\infty$ and the $\{p_i\}$ define a probability distribution.

The common feature of sequences of distributions in classes (Z) and (G) is the uniform convergence, for bounded t and large d, of the function $\prod_{i=1}^{d} (1 + p_{i,d} t)$ associated with the probabilities of the sets of distinct items towards a function $\lambda(t)$; sequences of

probability distributions in class (Z) are simply characterized by $\lambda(t)=e^t$. The uniform distribution is a special case of class (Z): $p_{i,d}=1/d$.

We shall work throughout this paper with two initial relations $R[X, Y]$ and $S[X, U]$. The distributions on the attributes Y and U are in class (Z) or in class (G) as indicated in the results of Sections 3.2 and 3.3. As for the join attribute, we shall assume that *the probability distribution on the join attribute X is uniform.*

2.3.3. Attributes common to two relations

The probability distribution on the domain of an attribute may be common to all the relations where this attribute is present, or the same attribute may have different distributions on different relations. The only condition is that the domain of an attribute is the same for all the relations where it appears, and that can easily be ensured.

For example, assume that we have two relations $R[X, Y]$ and $T[Y, U]$ and that the values that may appear in the attribute Y are, respectively, $\{y_1, y_2\}$ for R and $\{y_1, y_2, y_3\}$ for T. Assume also that the two possible values of Y that may appear in the relation R have the same probability 1/2 and that the three possible values of Y that may appear in the relation T have the same probability 1/3. Then we define the domain of the attribute Y as the set of all possible values in the two relations: $D_Y=\{y_1, y_2, y_3\}$ and we define two probability distributions on D_Y. The first distribution, denoted by p_R, is associated with R:

$$p_R(y_1)=p_R(y_2)=\tfrac{1}{2}, \quad p_R(y_3)=0.$$

The second distribution is associated with T and denoted by p_T:

$$p_T(y_1)=p_T(y_2)=p_T(y_3)=\tfrac{1}{3}.$$

In this paper, we partition the attributes of a given relation into the set of attributes on which it is joined to another relation and the set of attributes that are not part of the join. More precisely, we shall work with two relations $R[X, Y]$ and $S[X, U]$ to be joined on the (set of) attribute(s) X. Thus, an attribute common to both relations may appear in the sets of attributes Y and U, or be part of the equijoin or semijoin on X, and we may have to consider the case where this attribute has different distributions in R and in S. If the common attribute is not part of the join (first case), we shall define two probability distributions on Y and U as indicated above. The second case, where the common attribute is part of the join, will not really be considered here, as we assume that the probability distribution on X is uniform (but see [8, p. 590] for a preliminary result and Section 4.2 of this paper for a possible extension).

2.3.4. Limiting distributions

We are interested in the conditional limiting distribution of the equijoin or semijoin size, when the sizes r and s of the initial relations $R[X, Y]$ and $S[X, U]$ and the sizes of the domains D_X, D_Y and D_U grow large. This is equivalent to studying the probability distribution defined by the probability generating function

$f(x)=[y^r z^s]\Phi(x,y,z)/[y^r z^s]\Phi(1,y,z)$, where $\Phi(x,y,z)$ is itself a function associated with the sizes of the initial relations R and S and of their join. Let us stress here that in so far as the definition of $f(x)$ is concerned it does not matter if Φ is an ordinary or exponential generating function in the variables y or z, or if it is a counting or probability generating function, for either joint or conditional probabilities.

2.4. Generating functions

We present here the generating functions related to different schemes of relations and to the (projection or) join size. Variations on relation schemes or, equivalently, on balls and urns in the associated models, can be captured by the use of different generating functions, either ordinary (for distinguishable balls) or exponential (for undistinguishable balls) and for joint or conditional probabilities. We refer to [5] for a detailed study of the way to describe a relation (i.e. the set of all the legal instances) by a total generating function. We give below the function $\lambda(t)$ associated with the corresponding relation scheme or urn model, then the rules for the choice of the multivariate function $\Phi(x,y,z)$ and finally the functions Φ themselves. Of course, the choices for the functions λ and Φ are correlated and both depend on the schemes of the relations. From an "operational" point of view, the rules are such that the generating function Φ has the form φ^d; they can also be justified using ideas of combinatorial enumeration, as we shall indicate below.

2.4.1. Generating function associated with a relation or with an urn

In the urn model associated with the relation scheme $R[X, Y]$, the generating function $\lambda(t)=\sum_k a_k t^k$ describes the way balls are thrown into an urn. It also enumerates the possible sets of tuples with a common value on attribute X which may appear in an instance of relation R according to their cardinality.

- For the "classical" case (the balls are equivalent and the urns have unbounded capacity), the associated generating function describing what happens in an urn is the exponential function $\lambda(t)=e^t$. This corresponds to a relation $R[X, Y]$ with a key on attribute Y.
- If an urn can contain at most one ball we have $\lambda(t)=1+t$. In terms of a relation $R[X, Y]$, this corresponds to at most one tuple with a given value on the attribute X, i.e. this attribute is key of the relation R.
- If the relation $R[X, Y]$ has no key, the probability distribution on the domain of attribute Y has an influence on function λ. When this distribution is uniform, we use an enumerating function and choose $\lambda(t)=(1+t)^{d_Y}$; otherwise, we take $\lambda(t)=\prod_{i=1}^{d_Y}(1+p_{i,d_Y}t)$. The urn model we use here has d_Y types of balls and urns which can hold one ball of each type (this gives a global capacity of d_Y balls for each urn); the balls are distinguishable and each of them belongs to the ith type with a probability p_{i,d_Y}.

2.4.2. Choice of the function Φ

It has proved convenient to use the following transformation rule to get a function of the kind $\phi(x, y, z)^d$ and to emphasize the relationship between the joins and the urn models presented in Section 2.2:

If the relation $R[X, Y]$ has for key the attribute Y, we use a generating function for conditional probabilities, conditioned by the size of R and exponential in the variable marking the size of R. Similarly, if the relation $S[X, Z]$ has the attribute U for key, we use a generating function for conditional probabilities, conditioned by the size of the relation S and exponential in the variable marking the size of S.

For example, if R and S have, respectively, for keys Y and U, we use the generating function $\Phi(x, y, z) = \sum_{r,s,t} \text{Prob}(t/r, s)\, x^t (y^r/r!)\, (z^s/s!)$, where $\text{Prob}(t/r, s)$ is the probability that the join of R and S has size t, when the initial relations R and S have sizes r and s.

This rule allowed us in [7] to write the generating function for the projection under the general form (1) given in Section 2.2. We shall see presently that it is also useful for the study of the join sizes. It can be justified intuitively as follows. When relation R has attribute Y for key, it has at most d_Y tuples. The use of a conditional probability allows us to "forget" this upper bound $r \leqslant d_Y$ on the size r of the relation, and the division by $r!$ is equivalent to assuming that all the values of the attribute Y play the same role: the crucial point is not the exact set of values for Y, but the fact that these values are all distinct.

2.4.3. Generating functions for the joins

The generating functions for the join sizes can be expressed in terms of the generating functions λ_R and λ_S describing which sets of tuples with a given value on the join attribute X may appear in the relations $R[X, Y]$ and $S[X, U]$. In terms of the urn models A and B of Section 2.2, λ_R and λ_S describe how red balls and blue balls are allocated to any given urn.

Let us stress again that the result does not depend on our choosing an ordinary or exponential, counting or probability generating function: we just add a multiplicative factor independent of t in the term $[x^t y^r z^s]\Phi(x, y, z)$, which cancels in the conditional probability that the join has size t, knowing the sizes r and s of the initial relations R and S: $\text{Prob}(t/r, s) = [x^t y^r z^s]\Phi(x, y, z) / [y^r z^s]\Phi(1, y, z)$. Theorem 2.1 gives the generating functions for the semijoin and equijoin sizes when these functions are chosen according to the former rules. The notations are self-explanatory: $\text{Prob}(t, r, s)$ is the joint probability that the relations R and S have sizes r and s and that their join $R \bowtie S$ has size t; $\text{Prob}(t, r/s)$ is the probability that the relation R has size r and that the join $R \bowtie S$ has size t, knowing that the relation S has size s, etc.

Theorem 2.1. *Let $\Phi(x, y, z)$ be a generating function, where the variables y and z mark the sizes of the initial relations R and S, and the variable x marks the size of their semijoin or equijoin and with the following conventions for the choice of Φ.*

- *If each of the two relations R and S is either free or with a key X*:

$$\Phi(x,y,z)=\sum_{t,r,s}\text{Prob}(t,r,s)\,x^t y^r z^s.$$

- *If the attribute Y is key of the relation R and if the relation S is either free or has X for key*:

$$\Phi(x,y,z)=\sum_{t,r,s}\text{Prob}(t,s/r)\,x^t\frac{y^r}{r!}z^s.$$

- *If the relation R is free or has X for key and if the attribute U is key of the relation S*:

$$\Phi(x,y,z)=\sum_{t,r,s}\text{Prob}(t,r/s)\,x^t y^r\frac{z^s}{s!}.$$

- *If the attributes Y and U are keys, respectively, of the relations R and S*:

$$\Phi(x,y,z)=\sum_{t,r,s}\text{Prob}(t/r,s)\,x^t\frac{y^r}{r!}\frac{z^s}{s!}.$$

The function Φ is given by Fig. 6 for a semijoin, and by Fig. 7 for an equijoin. The terms a_k and b_l in Fig. 7 are the coefficients of λ_R and λ_S: $\lambda_R(t)=\sum_{k\geq 0}a_k t^k$ and $\lambda_S(t)=\sum_{l\geq 0}b_l t^l$.

Proof. We first "forget" some constant multiplicative factors to emphasize the structure of the tables. This is similar to what was done in the proof of Theorem 2 in [6, p. 229] and has no influence on the conditional distribution of the join size. The table of Fig. 6 is given in [6]. The ordinary counting generating functions for the equijoin are given in [8]. We then transform an ordinary counting generating

R	S	Φ	Asymptotic result
$X\dagger Y$	$X\dagger U$	$(\lambda_R(y)+\lambda_R(xy)[\lambda_S(z)-1])^{d_x}$	Theorem 3.1 and Corollary 3.5
$X\dagger Y$	$X\to U$	$(\lambda_R(y)+z\lambda_R(xy))^{d_x}$	[6, Theorem 4 and Corollary 4]
$X\dagger Y$	$U\to X$	$(\lambda_R(y)+\lambda_R(xy)[e^z-1])^{d_x}$	Theorem 3.1 and Corollary 3.4
$X\to Y$	$X\dagger U$	$(1+y+(1+xy)[\lambda_S(z)-1])^{d_x}$	[6, Theorem 3 and Corollary 2]
$X\to Y$	$X\to U$	$(1+y+z+xyz)^{d_x}$	[6, Theorem 5]
$X\to Y$	$U\to X$	$(1+y+(1+xy)[e^z-1])^{d_x}$	[6, Theorem 3 and Corollary 3]
$Y\to X$	$X\dagger U$	$(e^y+e^{xy}[\lambda_S(z)-1])^{d_x}$	Theorem 3.1 and Corollary 3.3
$Y\to X$	$X\to U$	$(e^y+ze^{xy})^{d_x}$	[6, Theorem 4 and Corollary 5]
$Y\to X$	$U\to X$	$(e^y+e^{xy}[e^z-1])^{d_x}$	Theorem 3.1 and Corollary 3.2

Fig. 6. Generating function for the sizes of the relations R, S and of their semijoin on the attribute X: $\{(x,y)\,|\,(x,y)\in R \text{ and } \exists z\colon (x,z)\in S\}$.

R	S	Φ	Asymptotic result
$X \dagger Y$	$X \dagger U$	$(\sum_{k,l\geq 0} a_k b_l x^{kl} y^k z^l)^{d_x}$	Theorem 3.6 and Corollary 3.9
$X \dagger Y$	$X \to U$	$(\lambda_R(y)+z\lambda_R(xy))^{d_x}$	[6, Theorem 4 and Corollary 4]
$X \dagger Y$	$U \to X$	$(\sum_{k,l\geq 0} a_k x^{kl} y^k \frac{z^l}{l!})^{d_x}$	Theorem 3.6 and Corollary 3.8
$X \to Y$	$X \dagger U$	$(\lambda_S(z)+y\lambda_S(xz))^{d_x}$	[6, Theorem 4 and Corollary 4]
$X \to Y$	$X \to U$	$(1+y+z+xyz)^{d_x}$	[6, Theorem 5]
$X \to Y$	$U \to X$	$(e^z+ye^{xz})^{d_x}$	[6, Theorem 4 and Corollary 5]
$Y \to X$	$X \dagger U$	$(\sum_{k,l\geq 0} b_k x^{kl} \frac{y^k}{k!} z^l)^{d_x}$	Theorem 3.6 and Corollary 3.8
$Y \to X$	$X \to U$	$(e^y+ze^{xy})^{d_x}$	[6, Theorem 4 and Corollary 5]
$Y \to X$	$U \to X$	$(\sum_{k,l\geq 0} x^{kl} \frac{y^k}{k!} \frac{z^l}{l!})^{d_x}$	Theorem 3.6 and Corollary 3.7

Fig. 7. Generating function for the sizes of the relations R, S and of their equijoin on the attribute X: $\{(x,y,z) \mid (x,y)\in R \text{ and } (x,z)\in S\}$. The coefficients a and b are defined by $\lambda_R(t)=\sum_{k\geq 0} a_k t^k$ and $\lambda_S(t)=\sum_{l\geq 0} b_l t^l$.

function into an exponential probability generating function when desired; straightforward computations give Fig. 7 (see again [6, proof of Theorem 2], for an example of such a computation).

A consequence of Theorem 2.1 worth noticing is that, when the attribute Y (or U) is key of the relation R (or S), the probability distribution on the domain of this attribute has absolutely no influence on the size of the joins: this distribution does not appear in the expression of Φ.

3. Asymptotic distributions

3.1. Presentation

In this section, we consider two initial relations R and S and their semijoin or equijoin on a common attribute X. The values taken by the two relations are assumed to be independent of each other, with the exception of Section 3.4. The relations R and S are each built on two attributes, $R[X, Y]$ and $S[X, U]$, respectively. We shall prove the asymptotic normality of the join size in several cases, and deduce from it that, as was shown in [6] for the projection, the probability distributions on the attributes Y and U have almost no importance.

The notion of convergence used in this paper is that of *convergence to a probability distribution* [4, p. 249]:

We say that a sequence of random variables (V_n) *converges to the probability distribution of a random variable* U, *when the sequence of distribution functions* (F_n)

associated with the U_n has for limit the distribution function F of the random variable U in every interval where F is continuous.

Sections 3.2 and 3.3 present two theorems (Theorems 3.1 and 3.6) relative to the asymptotic behavior of the urn models A and B when the parameter r, s (numbers of balls of each color) and d (number of urns) grow large, and when r and s are roughly proportional to d. In terms of conditional probabilities, these theorems give conditions ensuring that the distribution defined by $[y^r z^s]\varphi^d(x, y, z)/[y^r z^s]\varphi^d(1, y, z)$ converges to a probability distribution when $d, r, s \to +\infty$. Corollaries 3.2–3.5 and 3.7–3.9 present applications of these theorems to join sizes. Finally, Theorem 3.10 is relative to an extension of model B and the corresponding result for correlated relations is given in Corollary 3.11.

The use of general convergence results on urn models for estimating a join size deserve some attention. We shall always assume that the sizes r and s of the two relations to be joined and the size d_X of the domain of their join attribute are large, and that they are approximately proportional, but we also have to be aware of some additional constraints. For example, when the relation R has the attribute Y for key, assuming that the number r of its tuples grows large requires that the size of the domain D_Y becomes accordingly large (recall that $r \leqslant d_Y$), but the probability distribution on the attribute Y has no influence. When a relation is free, the domain size of its nonjoin attribute may either be fixed or grow large; if this is the case, we assume that the sizes d_Y or d_Z are independent of the size d_X of the join domain and independent of each other if both go to infinity.

We deduce from Theorem 2.1 that the relevant generating function $\Phi(x, y, z)$, where x marks the size of the semijoin, y the size of R and z the size of S or, equivalently, the generating function associated with model A of Section 2.2, has the general form

$$\Phi(x, y, z) = (\lambda_R(y) + \lambda_R(xy)(\lambda_S(z) - 1))^{d_X}.$$

In this formula, $\lambda_R(t) - 1$ is the generating function associated with the couples (x, y) of R whose x-value is fixed; we have already seen in Section 2.4 some examples of the functions λ_R corresponding to different schemes for the relation R and we recall them below.

- If R is free and the probability distribution on the attribute Y is given by $\{p_{i,d_Y}\}$, then $\lambda_R(t) = \prod_{i=1}^{d_Y}(1 + p_{i,d_Y} t)$.
- When X is the key of R, then $\lambda_R(t) = 1 + t$.
- If R has the attribute Y for key, we use the exponential function $\lambda_R(t) = \mathrm{e}^t$.

The function $\lambda_S(t)$ describes in a similar way the admissible sets of points in S. We study now what happens when d_Y, for example, grows large. If the relation R has the attribute Y for key, this has no influence on the function $\lambda_R(t) = \mathrm{e}^t$ (in order that the size of R goes to infinity, it is actually required that $d_Y \to +\infty$!). Similarly, if the join attribute X is key of the relation R, the variation of d_Y has no influence on the function λ_R and on the join size. Finally, if the relation R is free, we have to extend our theorems to take into account the fact that the generating function λ_R depends on d_Y; we use a sequence of functions λ_{R,d_Y}. As we deal with probability distributions on the domain

D_Y which are of type either (Z) or (G), this sequence converges uniformly towards a function $\lambda_{R,\infty}$, which we shall also denote by λ_R in the corollaries (see Section 2.3).

We shall assume in Theorems 3.1, 3.6 and 3.10 that each of the functions λ_R and λ_S is either equal to $1+t$ or satisfies the following property (we refer to it in the sequel as property $\mathscr{P}$), that holds for the exponential function and for functions associated with a free relation:

A function $\lambda(y)$ satisfies property $\mathscr{P}$ if it is entire, not affine, with positive coefficients, such that $\lambda(0)=1$, and such that there exists no entire function Λ and no integer $m\geqslant 2$ such that $\lambda(y)=\Lambda(y^m)$.

We are now ready to classify the joins according to the initial relation schemes. The semijoin is not a symmetrical operation: the semijoin of R and S is not equal to the semijoin of S and R. Assuming that each relation may independently have no key, or have for key any of its two attributes, there are nine possible cases for the semijoin of two relations. We can associate with several of these cases a generating function $\Phi(x,y,z)$ such that at least one of the coefficients $[y^r]\Phi$ or $[z^s]\Phi$ is easily computed; such cases were studied in our former paper [6]. The remaining cases are treated in this paper; the spirit of the proofs is the same, but the technical difficulties due to the form of the function Φ considerably lengthen the proof and justify a separate paper. We can sum up the situation for the semijoin as follows, according to the schemes of the relations R and S and choosing the generating function Φ according to Theorem 2.1.

- *R and S are free relations*: This is treated in Corollary 3.5, which is adapted from Theorem 3.1.
- *R is a free relation and S has X for key*: The computation of $[z^s]\Phi$ using the binomial theorem is easy. This was studied in [6, Theorem 4 and Corollary 5].
- *R is a free relation and S has U for key*: This is treated in this paper. See Corollary 3.4.
- *R has X for key and S is a free relation*: The computation of $[y^r]\Phi$ is easy; see [6, Theorem 3, Corollary 2].
- *R and S each have X for key*: The computation of $[y^r z^s]\Phi$ poses no difficulty. Here again, we have a gaussian limiting theorem [6, Theorem 5].
- *R has X for key and S has U for key*: See [6, Theorem 3 and Corollary 3].
- *R has Y for key and S is a free relation*: See Corollary 3.3.
- *R has Y for key and S has X for key*: See [6, Corollary 5 to Theorem 4].
- *R has Y for key and S has U for key*: This is a direct consequence of Theorem 3.1. See Corollary 3.2.

A similar classification holds for the equijoin. The generating function has the general form

$$\Phi(x,y,z)=\left(\sum_{k,l\geqslant 0} a_k b_l x^{kl} y^k z^l\right)^{d_x}.$$

We shall take advantage of the fact that the equijoin is a symmetrical operation to reduce the number of cases to be studied from nine to six. Furthermore, when one of

the initial relations has for key the join attribute X, the equijoin actually is a semijoin. This leads to the following classification for equijoins:

- *R and S are free relations*: This is Corollary 3.9.
- *R is a free relation and S has X for key*: The size of the equijoin is equal to the size of the semijoin of R with S, see [6, Theorem 4 and Corollary 4].
- *R is a free relation and S has U for key*: This is Corollary 3.8.
- *R and S each have X for key*: The size of the equijoin is equal to the size of either the semijoin of S with R or R with S, see [6, Theorem 5].
- *R has X for key and S has U for key*: The size of the equijoin is equal to that of the semijoin of S with R, see [6, Theorem 4 and Corollary 5].
- *R has Y for key and S has U for key*: See Corollary 3.7.

3.2. Semijoin sizes

We study in this section generating functions of the kind

$$\Phi(x, y, z) = (\lambda_R(y) + \lambda_R(xy)[\lambda_S(z) - 1])^d. \tag{2}$$

Theorem 3.1 is relative to the probability distribution defined by the probability generating function $f(x) = [y^r z^s]\Phi(x, y, z)/[y^r z^s]\Phi(1, y, z)$ when $d \to +\infty$ and r, s are proportional to d. The special cases where at least one of the functions $\lambda_R(t)$ or $\lambda_S(t)$ is equal to $1+t$ were studied in [6]; this corresponds to the cases where attribute X is key of at least one relation. We should point out that Theorem 3.1 could be extended to cover these cases; however, the corresponding theorems of [6] have less restrictive conditions on the relation sizes than Theorem 3.1 and this justifies a separate statement of these theorems.

We assume here that the functions λ_R and λ_S satisfy property $\mathscr{P}$. As a consequence, they are not affine functions, and none of the coefficients $[y^r]\Phi(x, y, z)$ or $[z^s]\Phi(x, y, z)$ can be extracted in the general case by an application of the binomial theorem. This has some implications on the proof techniques used to study the limit of the distribution defined by $f(x)$. We first give the general result (Theorem 3.1) for functions of the type (2), then applications relative to the distribution of the semijoin size in several cases (Corollaries 3.2–3.5). The proof of Theorem 3.1 is rather lengthy and we defer it until Section 5.3; we give after each corollary the ideas either to derive it from Theorem 3.1 or to modify the proof of Theorem 3.1 to obtain the desired result.

Theorem 3.1. *Let λ_R and λ_S be two functions satisfying property $\mathscr{P}$, and define $\Phi(x, y, z)$ by equation* (2). *Let $d, r, s \to +\infty$ in such a way that $r = Ad + o(d)$ and $s = Bd + o(d)$ for some strictly positive constants A and B. Suppose that A and B are such that the functions $g_R(y) = y\lambda_R'/\lambda_R(y)$ and $g_S(z) = z\lambda_S'/\lambda_S(z)$ satisfy $\lim_{y \to +\infty} g_R(y) > A$ and $\lim_{z \to +\infty} g_S(z) > B$. Then the probability distribution defined by the generating function $f(x) = [y^r z^s]\Phi(x, y, z)/[y^r z^s]\Phi(1, y, z)$ is asymptotically normal. The asymptotic mean*

and variance are $\mu = d\mu_0$ *and* $\sigma^2 = d\sigma_0^2$, *where* μ_0 *and* σ_0 *are constants defined in terms of the unique real positive solutions* ρ_R *and* ρ_S *of equations* $g_R(t) = A$ *and* $g_S(t) = B$:

$$\mu_0 = A(1 - 1/\lambda_S(\rho_S)),$$

$$\sigma_0^2 = (\rho_R g_R'(\rho_R) + A^2)\frac{\lambda_S(\rho_S) - 1}{\lambda_S^2(\rho_S)} - A^2\frac{\rho_S \lambda_S'^2(\rho_S)}{\lambda_S^4(\rho_S) g_S'(\rho_S)}.$$

In the applications of Theorem 3.1 to semijoin sizes, the limiting distributions are obtained for $r, s, d_X \to +\infty$; we indicate when the probability distributions on the attributes Y or U are also important for these distributions. In such cases, it is assumed that the domain sides d_Y and d_U grow to infinity independently of each other and of d_X. When the relation R is free and when $d_Y \to +\infty$, the function λ_R is the limit of the sequence of functions λ_{R,d_Y} (see Section 2.3); the function g_R is defined accordingly as $g_R(t) = t\lambda_R'/\lambda_R(t)$. A similar convention holds for the relation S. We also recall that, in Corollaries 3.2–3.5, we assume that the sizes r and s of the initial relations R and S satisfy $r = Ad_X + o(d_X)$ and $s = Bd_X + o(d_X)$, where A and B are strictly positive constants.

Corollary 3.2. *Let* $R[X, Y]$ *be a relation with key* Y, *and* $S[X, U]$ *a relation with key* U. *Then the probability distribution of the size of the semijoin of* R *and* S *on the attribute* X *is independent of the probability distributions on the domains* D_Y *and* D_U *of the key attributes* Y *and* U; *it is asymptotically normal, with asymptotic moments*

$$\mu = d_X\frac{e^B - 1}{e^B},$$

$$\sigma^2 = d_X\left(A\frac{e^B - 1}{e^{2B}} + A^2\frac{e^B - 1 - B}{e^{2B}}\right).$$

Proof. This is a simple instance of Theorem 3.1 in the case where $\lambda_R(t) = \lambda_S(t) = e^t$; hence, $g_R(t) = g_S(t) = t$. □

Corollary 3.3. *Let* $R[X, Y]$ *be a relation with a key on the attribute* Y *and* $S[X, U]$ *a free relation; the distribution on the attribute* U *is either in class* (Z) *or in class* (G). *Then the probability distribution on the attribute* Y *has no influence on the size of the semijoin. Assume that the constant* $B = \lim_{d_X, s \to +\infty} s/d_X$ *is chosen in such a way that* $\lim_{z \to +\infty} g_S(z) > B$. *Then the probability distribution of the size of the semijoin of* R *and* S *is asymptotically normal, with asymptotic moments*

$$\mu = d_X\left(1 - \frac{1}{\lambda_S(\rho_S)}\right),$$

$$\sigma^2 = d_X\left(A(A+1)\frac{\lambda_S(\rho_S) - 1}{\lambda_S^2(\rho_S)} - A^2\frac{\rho_S \lambda_S'^2(\rho_S)}{\lambda_S^4(\rho_S) g_S'(\rho_S)}\right).$$

Proof. According to Theorem 2.1, we choose for generating functions of the sizes of the initial relations and their semijoin $\Phi(x,y,z)=\sum_{r,s,t} p(t/r,s)x^t(y^r/r!)z^s = (e^y+e^{xy}(\lambda_S(z)-1))^d$. Theorem 3.1 applies with $\lambda_R(t)=e^t$ and $g_R(t)=t$. This gives a restriction on B but none on A. If $d_U \to +\infty$, then we adjust the proof of Theorem 3.1 to work with a sequence of functions indiced by the domain size $\lambda_{S,d_U}(t)=\prod_{i=1}^{d_U}(1+p_{i,d_U}t)$. This only requires to notice that we work in compact subsets of $]0, +\infty[$ or of the complex plane, and that the error terms which appear in the proof of Theorem 3.1 can still be chosen uniform in the necessary variables. Then the sequence of functions λ_{S,d_U} converges uniformly towards λ_S and we use $g_S(t)=t\lambda_S'/\lambda_S(t)$ to compute the variance of the limiting distribution. □

Corollary 3.4. *Let $R[X,Y]$ be a free relation and $S[X,U]$ a relation with key U. We assume that the constant $A=\lim_{d_X,r\to+\infty} r/d_X$ is such that $\lim_{y\to+\infty} g_R(y)>A$. The distribution on the domain of attribute Y is either in class (Z) or in class (G) and the distribution on the domain of attribute U has no influence. The semijoin size of R and S then converges asymptotically towards a normal distribution. The asymptotic moments are*

$$\mu = d_X \frac{e^B-1}{e^B},$$

$$\sigma^2 = d_X\left(\rho_R g_R'(\rho_R)\frac{e^B-1}{e^{2B}} + g_R^2(\rho_R)\frac{e^B-1-B}{e^{2B}}\right).$$

Proof. This result is symmetrical to Corollary 3.3: $\lambda_S(t)=e^t$ and $g_S(t)=t$; the constant A must satisfy the limiting condition $\lim_{y\to+\infty} g_R(y)>A$ and Theorem 3.1 applies, or can be adapted in the same way when $d_Y \to +\infty$. □

Corollary 3.5. *Let $R[X,Y]$ and $S[X,U]$ be two free relations. The constants $A=\lim_{d_X,r\to+\infty} r/d_X$ and $B=\lim_{d_X,s\to+\infty} s/d_X$ are assumed to satisfy $\lim_{y\to+\infty} g_R(y)>A$ and $\lim_{z\to+\infty} g_S(z)>B$. The probability distributions on the attributes Y and U belong to classes (Z) or (G). Then the distribution of the semijoin size converges towards a normal distribution; the asymptotic moments are $\mu=d_X\mu_0$ and $\sigma^2=d_X\sigma_0^2$ for suitable constants μ_0 and σ_0. As a special case, when the distributions on both attributes Y and U belong to class (Z), we have that $\mu_0 \approx (e^B-1)/e^B$ and $\sigma_0^2 \approx A(e^B-1)/e^{2B} + A^2(e^B-1-B)/e^{2B}$.*

Proof. When the domain sizes d_Y and d_U are fixed, Theorem 3.1 applies, with $\lambda_R(t)=\prod_{j=1}^{d_Y}(1+q_{j,d_Y}t)$ and $\lambda_S(t)=\prod_{k=1}^{d_U}(1+q'_{k,d_U}t)$. When d_Y or d_U grow large, we can adapt the proof of Theorem 3.1 to sequences of functions λ_{R,d_Y} and λ_{S,d_U} which are indexed, respectively, by d_Y and d_U and converge, respectively, towards functions λ_R and λ_S. If both distributions on domains D_Y and D_Z belong to class (Z), then $\lambda_R(t)=\lambda_S(t)=e^t$. □

Corollaries 3.2–3.5 give the same asymptotic moments for probability distributions of class (Z) on attributes Y and U. The meaning in terms of relations is as follows: The exact distributions on the attributes Y and U have no effect on the limiting distribution for free relations as long as they are not too far from uniform (i.e. in class Z); the exact relation schemes do not matter asymptotically as long as the join attribute is not a key of a relation.

3.3. Equijoin sizes

We shall assume in this part that none of the relations R or S has attribute X for key; this translates on the generating functions as $\lambda_R(t) \neq 1+t$ and $\lambda_S(t) \neq 1+t$. This is no restriction: when X is key of at least one of the initial relations, the equijoin actually has the same size as the semijoin of R with S or of S with R as noticed in Section 3.1. We first give a general theorem, namely Theorem 3.6, proved in Section 5.4, that applies to functions of the kind

$$\Phi(x,y,z)=\left(\sum_{k,l\geq 0} a_k b_l x^{kl} y^k z^l\right)^d. \tag{3}$$

We recall that we have $\lambda_R(y)=\sum_k a_k y^k$ and $\lambda_S(z)=\sum_l b_l z^l$. We then present corollaries relative to the distribution of equijoin sizes in several cases.

Theorem 3.6. *Let λ_R and λ_S be two functions satisfying Property $\mathscr{P}$, and define $\Phi(x,y,z)$ by equation* (3). *Let $d, r, s \to +\infty$ in such a way that $r=Ad+o(d)$ and $s=Bd+o(d)$ for some strictly positive constants A and B such that the functions $g_R(y)=y\lambda_R'/\lambda_R(y)$ and $g_S(z)=z\lambda_S'/\lambda_S(z)$ satisfy $\lim_{y\to+\infty} g_R(y)>A$ and $\lim_{z\to+\infty} g_S(z)>B$. Then the probability distribution defined by the generating function $f(x)=[y^r z^s]\Phi(x,y,z)/[y^r z^s]\Phi(1,y,z)$ is asymptotically normal with moments proportional to d. The asymptotic mean is $\mu=rs/d\approx ABd$. The asymptotic variance is defined in terms of the unique real positive solutions ρ_R and ρ_S of equations $g_R(t)=A$ and $g_S(t)=B$:*

$$\sigma^2 \approx d\rho_R g_R'(\rho_R)\rho_S g_S'(\rho_S).$$

Corollary 3.7. *Let $R[X,Y]$ and $S[X,U]$ be two relations with keys Y and U. Then the probability distributions on the attributes Y and U have no influence on the equijoin size. When the sizes r and s of the initial relations satisfy the assumptions of Theorem* 3.6, *the size of their equijoin is asymptotically normal. Its mean and its variance are $\mu=\sigma^2=rs/d_X\approx ABd_X$.*

Proof. This is simply Theorem 3.6 applied to the functions $\lambda_R(t)=\lambda_S(t)=e^t$. The distributions on the key attributes Y and U do not matter when the domain sizes d_Y and d_U are large enough so that r and s are of the order of d_X. □

Corollary 3.8. *Let $R[X, Y]$ be a free relation and $S[X, U]$ be a relation with key U. Then the size of the equijoin is independent of the probability distribution on the attribute U. When the sizes r and s of the initial relations satisfy the assumptions of Theorem* 3.6, *the size of their equijoin is asymptotically normal. Its mean is $\mu = rs/d_X$ and its asymptotic variance is $\sigma^2 \approx s\rho_R g'_R(\rho_R)$.*

Proof. The function $\lambda_S(t)$ is equal to e^t and the distribution on the domain D_U has no influence. We adapt the proof of Theorem 3.6 to work with a sequence of functions λ_{R,d_Y} as indicated in the proof of Corollary 3.3. The equijoin being a symmetrical operation, a similar corollary exists when Y is key of R and S is a free relation. □

Corollary 3.9. *Let $R[X, Y]$ and $S[X, U]$ be two free relations. When their sizes r and s satisfy the assumptions of Theorem* 3.6, *the size of their equijoin is asymptotically normal. The asymptotic mean is $\mu = rs/d_X$ and the asymptotic variance is $\sigma^2 = \sigma_0^2 d_X$ for a suitable constant $\sigma_0 > 0$. Furthermore, when the probability distributions on the domains of attributes Y and U are both in class (Z), the asymptotic variance is ABd_X.*

Proof. Apply Theorem 3.6 when the domain sizes d_Y and d_U are fixed, or extend its proof as indicated above when $d_Y, d_U \to +\infty$. □

3.4. Correlated relations

We study in this section the special case where the relations R and S to be joined are projections of an initial relation $T[X, Y, U]$. This means that the projections of R and S on their common attribute X are equal. This assumption has no interest for the semijoin: it simply means that the semijoin of R and S is always equal to R.

When the distribution on the attribute X is uniform, the generating function marking the sizes of the initial relations and of their equijoin is

$$\Phi(x, y, z) = \left(1 + \sum_{k,l \geq 1} a_k b_l x^{kl} y^k z^l\right)^{d_X}. \tag{4}$$

As in equation (3), the coefficients a_k and b_l define the functions λ_R and λ_S associated with the sets of tuples having a fixed value on the attribute X: $a_k = [y^k]\lambda_R(y)$ and $b_l = [z^l]\lambda_S(z)$.

The proof of equation (4) is in the same vein as that of formula (3) for independent relations. We first associate an urn model with the equijoin of the relations R and S. This model differs from model B only in the way to throw red and blue balls: at the end of the first phase, there is no urn with balls of only one color; each urn is either empty or contains at least one ball of each color. This comes from the fact that $\pi_X(R) = \pi_X(S)$: a possible value for the attribute X appears in both relations or not at all. We compute the generating function describing what can happen in any one urn, then take a sequence of d urns. Due to the fact that each urn is either empty or contains balls of the two initial colors (tuples of both initial relations), the generating

function associated with an urn and marking the numbers of balls of each color is simply $1+\sum_{k,l\geqslant 1} a_k b_l x^{kl} y^k z^l$; the final form (4) comes from the sequence of d_X urns.

When the relation R, for example, has the attribute X for key, the equijoin size $|R \bowtie S|$ has the same size as the relation S. This can easily be verified on the function Φ: $\lambda_R(t)=1+t$, which gives $a_0=a_1=1$ and $a_i=0$ for $i>1$; hence, $\Phi(x,y,z)=(1+\sum_{l\geqslant 1} b_l x^l y z^l)^{d_X}=(1+y(\lambda_S(xz)-1))^{d_X}$.

We give Theorem 3.10 for functions of the form (4). As an application, Corollary 3.11 shows that the size of the equijoin of two relations obtained by projections of an initial relation is once again asymptotically normal under suitable conditions. Actually, the important condition is not the existence of a common relation $T[X, Y, U]$ whose projections are joined, but the special form of the generating function of the sizes, by the assumption that $\pi_X(R)=\pi_X(S)$ for all admissible couples (R, S).

Theorem 3.10. *Let $\lambda_R(t)=\sum_k a_k t^k$ and $\lambda_S(t)=\sum_l b_l t^l$ be two functions satisfying property $\mathscr{P}$ and define $\Phi(x,y,z)=(1+\sum_{k,l\geqslant 1} a_k b_l x^{kl} y^k z^l)^d$. Let $d, r, s \to +\infty$ in such a way that $r= Ad+o(d)$ and $s=Bd+o(d)$, for some strictly positive constants A and B. Suppose that A and B are such that the functions $g_R(y)=y\lambda_R'/\lambda_R(y)$ and $g_S(z)=z\lambda_S'/\lambda_S(z)$ satisfy $\lim_{y\to+\infty} g_R(y)>A$ and $\lim_{z\to+\infty} g_S(z)>B$. Then the probability distribution defined by the generating function $f(x)=[y^r z^s]\Phi(x,y,z)/[y^r z^s]\Phi(1,y,z)$ is asymptotically normal. The asymptotic mean and variance are $\mu=d\mu_0$ and $\sigma^2=d\sigma_0^2$ for suitable constants μ_0 and σ_0.*

Corollary 3.11. *Let R and S be two relations with a common attribute X, such that, for every instance of the couple (R, S), we have $\pi_X(R)=\pi_X(S)$. We assume that the sizes r and s of the two relations satisfy the hypothesis of Theorem* 3.10 *and that none of the relations R or S has the attribute X for key. Then the probability distribution of the size of the equijoin of R and S on their common attribute X, conditioned by the sizes of the initial relations, is asymptotically normal when $d_X, r, s \to +\infty$, with mean and variance proportional to d_X.*

Proof. Corollary 3.11 is a straightforward application of Theorem 3.10 when the relations R and S have for keys, respectively, the attribute Y and the attribute U. When the relation R is free and the relation S has the attribute U for key, Theorem 3.10 applies without modification if the size d_Y of the nonjoin attribute is fixed. If d_Y also grows to infinity, we extend the proof of Theorem 3.10 in the usual way. The cases where the relation S is free are similar. □

4. Discussion and conclusion

4.1. Summary of our results

We have associated a class of generating functions with each operator of the relational algebra whose effect on the relation sizes is not trivial: the projection, the

semijoin or the equijoin. The effect of the other operators of the relational algebra on the relation sizes is straightforward: the size of the cartesian product $R \times S$ is equal to the product of the sizes of R and S; the intersection $R \cap S$ is closely related to the join of two relations that have the join attribute X for key and has the same size; the size of a relation obtained by a set operation can be deduced from the sizes of the initial relations and from the size of the relation obtained by any other set operation; the selection of the tuples of a relation R satisfying a boolean condition C can be seen as the intersection of R and of the set S of all tuples satisfying the condition C (of course, we still have to compute the size of $S \ldots$).

We have related the projection sizes and join sizes to a well-known probabilistic problem, i.e. an urn model, or to extensions of it. We have then proved that the size of a relation obtained by a projection or a join often follows asymptotically a normal distribution. Our results are in the spirit of asymptotic studies of the "classical" occupancy problem in urn models [11, 12].

Query optimizers often use very crude assumptions; for example, in System R, the size of a relation obtained by selecting the tuples satisfying same condition is assumed to be a constant fraction of the size of the initial relation. As for the join size, most of the studies on the subject limit themselves to the average size (see e.g. [15]). The present paper proves mathematically that, under a wide range of conditions, *the distribution of the join size is "almost" normally distributed*, with a rather small variance. As a consequence, there are few deviations from the mean value and it is often enough for practical purposes to use only this average relation size. We have also studied the distributions on the non-join attributes: *either they have no influence on the asymptotic result or they influence only slightly the average value and variance of the limiting distribution*, without changing the type of this distribution, which remains gaussian.

The single most important assumption required to have a normal limiting distribution may well be the independence of the urns, or equivalently, of the sets of tuples having a common value on the join attribute. From a mathematical point of view, this means that the generating functions we use are *products* (or powers) of simpler functions.

4.2. Possible extensions

The work presented in this paper can be extended to allow weaker assumptions on the database model. An obvious extension would be to allow for a nonuniform distribution on the join attribute X. This corresponds to using a different function φ_i for each value of the domain D_X and can be simply modelized. For instance, the general form of the function describing the sizes of two relations and of their equijoin is, in the case of a uniform distribution on the join attribute,

$$\Phi(x, y, z) = (\lambda_R(y) + \lambda_R(xy)(\lambda_S(z) - 1))^{d_x}$$

(see Fig. 6). If the distribution on the join attribute is given by $\{p_i, 1 \leqslant i \leqslant d_X\}$, then the function associated with the sizes of the initial relations and of their join becomes

$$\Phi(x, y, z) = \prod_{i=1}^{d_X} (\lambda_R(p_i y) + \lambda_R(p_i xy)(\lambda_S(p_i z) - 1)).$$

In the same vein if the distributions on the domain of the join attribute X are different for the two relations R and S, a slight modification of the generating function can take care of it. Again this means working with a generating function $\prod_{i=1}^{d} \varphi_i$ instead of φ^d. Finally, if the attributes of a relation are not independent (we recall again that the existence of a key is not contradictory with the independence of attributes), some relationship can still be captured by a suitable generating function. For instance, consider a relation R such that the set of values of the attribute Y associated with a given value of the attribute X, or with a given urn, differs according to the urn; then we shall use a different function $\lambda_R(t)$ to describe what may happen in each urn. This amounts once more to using a generating function of the kind $\prod_{i=1}^{d} \varphi_i$ instead of φ^d.

From a mathematical point of view, these extensions all have in common the fact that they deal with a product of functions, whose number of terms grows large, instead of a "large" power of a function. Another, probably simpler, extension might be to extend the range of applicability of our results to allow for a nonproportional growth of the sizes of the initial relations and the domain size (r, s and d_X in this paper). By analogy with the classical urn model (see [11, Ch. 6]), and using recent results on asymptotic distributions [3], it should be possible to obtain results of asymptotic normality.

5. Proofs of theorems

5.1. Notations and preliminary results

Our assumptions for all the theorems in this section are as follows: $r = Ad + o(d)$ and $s = Bd + o(d)$, with A and B strictly positive constants. The functions Φ that we shall study are defined from two initial functions λ_R or λ_S (both satisfying property $\mathscr{P}$) by one of the equations (2), (3) or (4). We define $g_R(y) = y\lambda'_R(y)/\lambda_R(y)$ and $g_S(z) = z\lambda'_S(z)/\lambda_S(z)$. We recall here some properties that we shall need to prove our theorems (see [6] for the proofs[2]):

Lemma 5.1. *Let $\lambda(y)$ be a function satisfying property $\mathscr{P}$ and define $g(y) = y\lambda'(y)/\lambda(y)$. Then g is increasing on the interval $[0, +\infty[$.*

[2] Lemma 5.1 and Lemma 5.2 are, respectively, Lemma A and Lemma C in [6].

We use Lemma 5.1 as follows: As $g(0)=0$ and as g is increasing on $[0,+\infty[$, the equation $g(y)=A$ has at most one unique real positive solution ρ_0 when A is a positive real number; this solution exists if and only if

$$\lim_{y\to+\infty} g(y)>A.$$

Lemma 5.2. *Let λ be a function satisfying property $\mathscr{P}$ and let $\alpha\in]0,\pi[$. Let y vary in a compact subset of $]0,+\infty[$. Then there exists a constant $C>0$ independent of α such that, for all θ satisfying $\alpha<|\theta|<\pi$, and for all y in the compact subset, the following bound holds:*

$$|\lambda(ye^{i\theta})|\leqslant\lambda(y)(1-C\alpha^2).$$

5.2. Sketch of the proofs

We present here the general method that we shall use for proving Theorems 3.1, 3.6 and 3.10; the detailed proofs will be found in Sections 5.3–5.5. We study the conditional probability distribution defined from a function $\Phi(x,y,z)=\varphi^d(x,y,z)$ by the generating function $f(x)=[y^r z^s]\Phi(x,y,z)/[y^r z^s]\Phi(1,y,z)$, when $r,s,d\to+\infty$. We first evaluate $\psi(x)=[y^r z^s]\Phi(x,y,z)$ for x real in $]0,1[$ and close to 1. In the cases we study here, $\psi(x)$ cannot be computed by applying the binomial theorem and we have to apply Cauchy's formula twice (see e.g. [1, pp. 135–137]). To this effect, we first consider Φ as an analytic function of y; the variables x and z are then parameters. We apply Cauchy's formula to obtain $[y^r]\Phi(x,y,z)$

$$[y^r]\Phi(x,y,z)=\frac{1}{2i\pi}\oint\Phi(x,y,z)\frac{dy}{y^{r+1}}.$$

This function is again an analytic function of the variable z, with parameter x and Cauchy's formula applied once more gives

$$\psi(x)=\frac{1}{(2i\pi)^2}\oint\oint\Phi(x,y,z)\frac{dy}{y^{r+1}}\frac{dz}{z^{s+1}},$$

with integration paths around the origin. As the function Φ is analytic and has an infinite radius of convergence separately in y and in z,[3] we choose for integration path in each case a circle passing through a *saddle point*: It is centered at the origin, and its radius is defined by an equation of the kind $h'_y=0$ or $h'_z=0$, with $h(x,y,z)=\log\Phi(x,y,z)-(r+1)\log y-(s+1)\log z$.

Once we have an approximation of $\psi(x)$ for $x\to1^-$, we show the pointwise convergence towards $e^{t^2/2}$ of the normalized Laplace transform $e^{t\mu/\sigma}f(e^{-t/\sigma})=e^{t\mu/\sigma}\psi(e^{-t/\sigma})/\psi(1)$, for suitably chosen values of μ and σ and for any fixed t in the

[3] This may not be true for all values of x in Theorems 3.6 and 3.10; however, this holds for $x\in]0,1]$, which is the range we are interested in.

interval $[0,+\infty[$. We then conclude the convergence of the probability distribution defined by $f(x)$ towards a normal distribution of mean μ and variance σ^2.

This type of proof comprises two parts: approximation of a probability generating function $f(x)$ by analytical techniques, and then the application of a limiting theorem from general probability theory. It was already used in [6]. However, the first part was simpler due to the fact that at least one of the coefficients $[y^r]\Phi$ or $[z^s]\Phi$ that appear in the computation of $f(x)$ could easily be obtained and that $f(x)$ was then approximated with just one application of Cauchy's formula.

5.3. *Semijoin sizes and Theorem 3.1*

We recall that the generating function $\Phi(x,y,z)$ is

$$\Phi(x,y,z)=(\lambda_R(y)+\lambda_R(xy)(\lambda_S(z)-1))^d,$$

with x marking the semijoin size, y the size of the relation R and z the size of the relation S. We want to compute an approximation of

$$\psi(x)=\frac{1}{(2i\pi)^2}\oint\oint e^{h(x,y,z)}\,dy\,dz, \tag{5}$$

with

$$h(x,y,z)=d\log[\lambda_R(y)+\lambda_R(xy)(\lambda_S(z)-1)]-(r+1)\log y-(s+1)\log z$$

and for x real and close to 1. We shall use the following identity on the function h to simplify some computations:

$$h(1,y,z)=\{d\log\lambda_R(y)-(r+1)\log y\}+\{d\log\lambda_S(z)-(s+1)\log z\}.$$

5.3.1. *Choice of integration paths*

The integration contours in equation (5) are closed curves around the origin; we choose two circles centered at the origin and with respective radii $y(x)$ and $z(x)$ defined by

$$\frac{\partial h}{\partial y}(x,y,z)=0,$$

$$\frac{\partial h}{\partial z}(x,y,z)=0.$$

We shall first solve this system for $x=1$, then get an approximate solution for $x=1+\varepsilon$.

For $x=1$, $h(1,y,z)=\{d\log\lambda_R(y)-(r+1)\log y\}+\{d\log\lambda_S(z)-(s+1)\log z\}$, and the system defining $y(1)$ and $z(1)$ can be rewritten to obtain one equation in y and one equation in z:

$$y\frac{\lambda_R'}{\lambda_R}(y)\quad(=g_R(y))=\frac{r+1}{d}, \tag{6}$$

$$z\frac{\lambda_S'}{\lambda_S}(z)\quad(=g_S(z))=\frac{s+1}{d}. \tag{7}$$

The terms $(r+1)/d$ and $(s+1)/d$ have A and B for limits when $d, r, s \to +\infty$. By Lemma 5.1, equations (6) and (7) have unique real positive solutions if and only if

$$\lim_{y \to +\infty} g_R(y) > A, \tag{8}$$

$$\lim_{z \to +\infty} g_S(z) > B. \tag{9}$$

We now assume that these conditions are satisfied. Let (ρ_R, ρ_S) be the unique solution of the system (6), (7) among real positive numbers. Although the terms ρ_R and ρ_S depend on r, s and d, they can be restricted to compact neighborhoods in $]0, +\infty[$, respectively, of $g_R^{-1}(A)$ and of $g_S^{-1}(B)$. For $x = 1+\varepsilon$, we solve the initial system, which we first rewrite as

$$y \frac{\lambda_R'(y) + x\lambda_R'(xy)(\lambda_S(z)-1)}{\lambda_R(y) + \lambda_R(xy)(\lambda_S(z)-1)} = g_R(\rho_R),$$

$$z \frac{\lambda_R(xy)\lambda_S'(z)}{\lambda_R(y) + \lambda_R(xy)(\lambda_S(z)-1)} = g_S(\rho_S).$$

We look for approximate solutions of the kind $y(x) = \rho_R(1+v)$ and $z(x) = \rho_S(1+u)$. The functions $\lambda_R, \lambda_S, \lambda_R'$ and λ_S' can be expanded near the points ρ_R and ρ_S. For example,

$$\lambda_R(y) = \lambda_R(\rho_R) + (y-\rho_R)\lambda_R'(\rho_R) + O((y-\rho_R)^2 \| \lambda_R'' \|),$$

with an error term uniform in r, s and d: $O((y-\rho_R)^2 \| \lambda_R'' \|) = O((y-\rho_R)^2)$. This gives the approximate system

$$\left(1 - \frac{1}{\lambda_S(\rho_S)}\right)\varepsilon + v + O(\varepsilon^2) + O(v^2) + O(vu) + O(\varepsilon u) = 0,$$

$$\frac{g_R(\rho_R) g_S(\rho_S)}{\lambda_S(\rho_S)} \varepsilon + \rho_S g_S'(\rho_S) u + O(\varepsilon^2) + O(v^2) + O(u^2) + O(vu) + O(\varepsilon u) = 0.$$

Solving, we get

$$v = -\alpha_1 \varepsilon + O(\varepsilon^2), \tag{10}$$

$$u = -\alpha_2 \varepsilon + O(\varepsilon^2). \tag{11}$$

The coefficients α_1 and α_2 are strictly positive:

$$\alpha_1 = 1 - \frac{1}{\lambda_S(\rho_S)},$$

$$\alpha_2 = g_R(\rho_R) \frac{\lambda_S'(\rho_S)}{g_S'(\rho_S)\lambda_S^2(\rho_S)}.$$

We give here the values of some derivatives of h at the point $(1, \rho_R, \rho_S)$ that we shall need below:

$$h'_x(1, \rho_R, \rho_S) = d g_R(\rho_R) \frac{\lambda_S(\rho_S) - 1}{\lambda_S(\rho_S)}.$$

The functions h'_y, h'_z and h''_{yz} are equal to 0 at the point $(1, \rho_R, \rho_S)$; other derivatives of second order of h are

$$h''_{x^2}(1, \rho_R, \rho_S) = d \frac{\lambda_S(\rho_S) - 1}{\lambda_S(\rho_S)} \left(\rho_R g'_R(\rho_R) - g_R(\rho_R) + \frac{g_R^2(\rho_R)}{\lambda_S(\rho_S)} \right),$$

$$h''_{y^2}(1, \rho_R, \rho_S) = d \frac{g'_R(\rho_R)}{\rho_R},$$

$$h''_{z^2}(1, \rho_R, \rho_S) = d \frac{g'_S(\rho_S)}{\rho_S},$$

$$h''_{xy}(1, \rho_R, \rho_S) = d g'_R(\rho_R)(1 - 1/\lambda_S(\rho_S)),$$

$$h''_{xz}(1, \rho_R, \rho_S) = d g_R(\rho_R) \frac{\lambda'_S}{\lambda_S^2}(\rho_S).$$

Near the point $(1, \rho_R, \rho_S)$ we also have $\| h''' \| = O(d)$.

5.3.2. Evaluation of $\psi(x)$

We recall that

$$\psi(x) = \frac{1}{(2i\pi)^2} \oint\oint e^{h(x,y,z)} \, dy \, dz.$$

The integration paths are two circles $y = y(x)e^{i\theta}$ and $z = z(x)e^{i\tau}$, centered at the origin and whose radii were determined in Section 5.3.1: for $x = 1 + \varepsilon$, we take $y(x) = \rho_R(1 + v) = \rho_R(1 - \alpha_1 \varepsilon + O(\varepsilon^2))$ and $z(x) = \rho_S(1 + u) = \rho_S(1 - \alpha_2 \varepsilon + O(\varepsilon^2))$. We choose here x fixed near 1, real and smaller than 1. This fixes $y(x)$ and $z(x)$ near ρ_R and ρ_S. Let $\alpha \in]0, \pi/2[$; we divide the integral giving $\psi(x)$ in two parts: $\psi(x) = I_1 + I_2$, with I_1 corresponding to the integral on $\mathscr{E} =]-\alpha, +\alpha[^2$ and I_2 to the complementary part $[-\pi, +\pi]^2 \setminus \mathscr{E}$. We can approximate I_1 and show that I_2 is exponentially negligible for a suitable choice of α.

Approximation of I_1. I_1 is obtained by integration on arcs $y = y(x)e^{i\theta}$ and $z = z(x)e^{i\tau}$ for $|\theta|, |\tau| < \alpha$:

$$I_1 = \frac{1}{(2i\pi)^2} \int_{\mathscr{E}} e^{h(x,y,z)} \, dy dz,$$

$$= \frac{e^{h(x,y(x),z(x))}}{(2i\pi)^2} \int_{\mathscr{E}} e^{h(x, y(x)e^{i\theta}, z(x)e^{i\tau}) - h(x,y(x),z(x))} \, dy dz.$$

For y near $y(x)$ and z near $z(x)$ and as $\|h'''\| = O(d)$ in $\mathscr{E}$, we have

$$h(x,y,z) = h(x,y(x),z(x)) + (y-y(x))h'_y + (z-z(x))h'_z + \tfrac{1}{2}(y-y(x))^2 h''_{y^2}$$
$$+ \tfrac{1}{2}(z-z(x))^2 h''_{z^2} + (y-y(x))(z-z(x))h''_{yz}$$
$$+ O(d(y-y(x))^3) + O(d(z-z(x))^3).$$

The derivatives of h in this formula are taken at the point $(x, y(x), z(x))$. We recall that, by definition of $y(x)$ and $z(x)$, we have $h'_y(x,y(x),z(x)) = h'_z(x,y(x),z(x)) = 0$. Let h_i be any one of the derivatives of second order:

$$h_i(x,y(x),z(x)) = h_i(1,\rho_R,\rho_S)(1+O(x-1)).$$

As $h''_{y^2}(1,\rho_R,\rho_S)$ and $h''_{z^2}(1,\rho_R,\rho_S)$ are both strictly positive, so are $h''_{y^2}(x,y(x),z(x))$ and $h''_{z^2}(x,y(x),z(x))$ for x close to 1. We have

$$I_1 = e^{h(x,y(x),z(x))}/(2i\pi)^2 \int_{\mathscr{E}} \exp(\tfrac{1}{2}(y-y(x))^2 h''_{y^2} + \tfrac{1}{2}(z-z(x))^2 h''_{z^2}$$
$$+ (y-y(x))(z-z(x))h''_{yz})(1+\delta(\theta,\tau))\,dy\,dz,$$

with an error term $\delta(\theta,\tau) = O(d\theta^3) + O(d\tau^3)$. On the set $\mathscr{E} =]-\alpha, +\alpha[^2$, $\delta(\theta,\tau)$ has an upper bound $\delta(\alpha,\alpha) = O(d\alpha^3)$ that is independent of the integrand. Define

$$I'_1 = \int_{\mathscr{E}} \exp(\tfrac{1}{2}(y-y(x))^2 h''_{y^2} + \tfrac{1}{2}(z-z(x))^2 h''_{z^2} + (y-y(x))(z-z(x))h''_{yz})\,dy\,dz.$$

We have $I_1 = (e^{h(x,y(x),z(x))}/(2i\pi)^2)\,(1+\delta(\alpha,\alpha))\,I'_1$ and the integral I'_1 can be computed approximately.

To simplify the notations, let us write y_0 and z_0 for $y(x)$ and $z(x)$ in the next paragraphs. We change the two integration paths as follows: The integration path $\{|\theta| \leqslant \alpha\}$ becomes $\gamma_{1,y} \cup \gamma'_{1,y} \cup \gamma_{2,y}$, with $\gamma_{1,y} = \{y_0(1-v) - iy_0 \sin\alpha,\ 0 \leqslant v \leqslant 1-\cos\alpha\}$, $\gamma'_{1,y} = \{y_0(1-v) + iy_0 \sin\alpha,\ 0 \leqslant v \leqslant 1-\cos\alpha\}$ and $\gamma_{2,y} = \{y_0 + iy_0 v,\ -\sin\alpha \leqslant v \leqslant +\sin\alpha\}$. The integration path for τ is modified in a similar way and becomes $\gamma_{1,z} \cup \gamma'_{1,z} \cup \gamma_{2,z}$ (just substitute z_0 for y_0). We divide the integral I'_1 into nine parts and show that each one of them can be neglected, with the exception of the integral on $\gamma_{2,y} \times \gamma_{2,z}$.

- On $\gamma_{1,y} \times \gamma_{1,z}$: $y = y_0(1-v) - iy_0 \sin\alpha$ and $z = z_0(1-t) - iz_0 \sin\alpha$, with $v, t \in [0, 1-\cos\alpha]$. We seek an upper bound on the integral

$$y_0 z_0 \int_{[0,\,1-\cos\alpha]^2} \exp\left(\frac{y_0^2}{2}(v+i\sin\alpha)^2\, h''_{y^2} + \frac{z_0^2}{2}(t+i\sin\alpha)^2\, h''_{z^2}\right.$$
$$\left. + y_0 z_0 (v+i\sin\alpha)(t+i\sin\alpha) h''_{yz}\right) dv\,dt.$$

Its modulus is bounded by

$$y_0 z_0 \int_{[0,\,1-\cos\alpha]^2} \exp\left(\Re\left\{ \frac{y_0^2}{2}(v+\mathrm{i}\sin\alpha)^2\, h''_{y^2} + \frac{z_0^2}{2}(t+\mathrm{i}\sin\alpha)^2\, h''_{z^2} + y_0 z_0 (v+\mathrm{i}\sin\alpha)(t+\mathrm{i}\sin\alpha) h''_{yz} \right\} \right) \mathrm{d}v\,\mathrm{d}t.$$

The derivatives of h are real and positive at point (x, y_0, z_0) and we have for upper bound

$$y_0 z_0 \int_{[0,\,1-\cos\alpha]^2} \exp\left(\tfrac{1}{2}\left(y_0^2 h''_{y^2}(v^2-\sin^2\alpha) + 2y_0 z_0 h''_{yz}(vt-\sin^2\alpha) + z_0^2 h''_{z^2}(t^2-\sin^2\alpha)\right)\right) \mathrm{d}v\,\mathrm{d}t.$$

Now $v^2-\sin^2\alpha \leqslant 2\cos\alpha\,(\cos\alpha-1) \leqslant -\alpha^2/2$; a similar upper bound holds for the terms in t^2 and vt. The integral has for upper bound $y_0 z_0 (1-\cos\alpha)^2 \mathrm{e}^{-\alpha^2 C/4}$, with a factor $C = y_0^2 h''_{y^2} + 2y_0 z_0 h''_{z^2} + z_0^2 h''_{z^2}$ that is of order d: $C = 4dC_0(1+o(1))$ for a constant $C_0 > 0$. As y_0 and z_0 are in a compact subset of $]0, +\infty[$ when x is close to 1, the integral on $\gamma_{1,y} \times \gamma'_{1,z}$ is finally $\mathrm{O}(\alpha^4 \mathrm{e}^{-C_0 d\alpha^2})$.

- On $\gamma_{1,y} \times \gamma_{2,z}$, $y = y_0(1-v) - \mathrm{i}y_0 \sin\alpha$ and $z = z_0 + \mathrm{i}z_0 t$, with $v \in [0, 1-\cos\alpha]$ and $t \in [-\sin\alpha, +\sin\alpha]$. The integral is then

$$-\mathrm{i}y_0 z_0 \int_0^{1-\cos\alpha} \int_{-\sin\alpha}^{+\sin\alpha} \exp\left(\frac{y_0^2}{2}(v+\mathrm{i}\sin\alpha)^2 h''_{y^2} - \mathrm{i}y_0 z_0 t (v+\mathrm{i}\sin\alpha) h''_{z^2} - \frac{z_0^2}{2} t^2 h''_{z^2} \right) \mathrm{d}v\,\mathrm{d}t.$$

Again its modulus is bounded by

$$y_0 z_0 \int_0^{1-\cos\alpha} \int_{-\sin\alpha}^{+\sin\alpha} \exp\left(\frac{y_0^2}{2}(v^2-\sin^2\alpha) h''_{y^2} + y_0 z_0 t h''_{yz} \sin\alpha - \frac{z_0^2}{2} t^2 h''_{z^2} \right) \mathrm{d}v\,\mathrm{d}t.$$

The exponent $(y_0^2/2)(v^2-\sin^2\alpha)h''_{y^2} + y_0 z_0 t h''_{yz} \sin\alpha - (z_0^2/2)t^2 h''_{z^2}$ is bounded for $v \in [0, 1-\cos\alpha]$ and $t \in [-\sin\alpha, +\sin\alpha]$ by $(-(y_0^2/4)h''_{y^2} + y_0 z_0 h''_{yz})\alpha^2$. The modulus of the integral is then less than $2y_0 z_0 (1-\cos\alpha)\sin\alpha\, \mathrm{e}^{-(y_0^2 h''_{y^2}/4 - y_0 z_0 h''_{yz})\alpha^2}$. We next take advantage of the fact that h''_{y^2} is of order exactly d, while h''_{yz} is $\mathrm{o}(d)$, and we deduce from it that the integral on $\gamma_{1,y} \times \gamma_{2,z}$ is $\mathrm{O}(\alpha^3 \mathrm{e}^{-C_1 d\alpha^2})$ for some constant $C_1 > 0$.

- By symmetry, the integrals on $\gamma_{1,y} \times \gamma'_{1,z}, \gamma'_{1,y} \times \gamma_{1,z}$ and $\gamma'_{1,y} \times \gamma'_{1,z}$ are $\mathrm{O}(\alpha^4 \mathrm{e}^{-C_0 d\alpha^2})$ and the integrals on $\gamma_{2,y} \times \gamma_{1,z}, \gamma'_{1,y} \times \gamma_{2,z}$ and $\gamma_{2,y} \times \gamma'_{1,z}$ are $\mathrm{O}(\alpha^3 \mathrm{e}^{-C_1 d\alpha^2})$.

- On $\gamma_{2,y}\times\gamma_{2,z}$: $y=y_0+iy_0v$ and $z=z_0+iz_0t$, with $|v|$, $|t|\leqslant\sin\alpha$. We have to compute the integral

$$-y_0z_0\int_{[-\sin\alpha,\,+\sin\alpha]^2}\exp(-\tfrac{1}{2}(v^2y_0^2h''_{y^2}+2vty_0z_0h''_{yz}+t^2z_0^2h''_{z^2}))\,\mathrm{d}v\,\mathrm{d}t.$$

A rescaling of the variables $v\mapsto y_0v$ and $t\mapsto z_0t$ gives a simpler form

$$-\int_{[-y_0\sin\alpha,\,+y_0\sin\alpha]\times[-z_0\sin\alpha,\,+z_0\sin\alpha]}\exp(-\tfrac{1}{2}(v^2h''_{y^2}+2vth''_{yz}+t^2h''_{z^2}))\,\mathrm{d}v\,\mathrm{d}t.$$

We now have to compute an integral of the type

$$J(a,b,c)=\int_{[-y_0\sin\alpha,\,+y_0\sin\alpha]\times[-z_0\sin\alpha,\,+z_0\sin\alpha]}\exp(-\tfrac{1}{2}(av^2+2bvt+ct^2))\,\mathrm{d}v\,\mathrm{d}t$$

for real positive a and c and real b; the integral on $\gamma_{2,y}\times\gamma_{2,z}$ is $-J(h''_{y^2},h''_{yz},h''_{z^2})$. We first integrate with respect to t:

$$\int_{-z_0\sin\alpha}^{+z_0\sin\alpha}\exp(-\tfrac{1}{2}(av^2+2bvt+ct^2))\,\mathrm{d}t$$

$$=\exp\left(-\frac{ac-b^2}{2c}v^2\right)\int_{-z_0\sin\alpha}^{+z_0\sin\alpha}\exp\left(-\frac{c}{2}\left(t-\frac{bv}{c}\right)^2\right)\mathrm{d}t.$$

The substitution of $u=\sqrt{c}\,(t-bv/c)$ for the integration variable allows us to compute this last integral

$$\int_{-z_0\sin\alpha}^{+z_0\sin\alpha}\exp\left(-\frac{c}{2}\left(t-\frac{bv}{c}\right)^2\right)\mathrm{d}t=\frac{1}{\sqrt{c}}\int_{\sqrt{c}(-z_0\sin\alpha-(bv/c))}^{\sqrt{c}(+z_0\sin\alpha-(bv/c))}\mathrm{e}^{-(u^2/2)}\,\mathrm{d}u$$

$$=\frac{1}{\sqrt{c}}\left(\int_{-\infty}^{+\infty}\mathrm{e}^{-(u^2/2)}\,\mathrm{d}u\right.$$

$$+\mathrm{O}\left(\exp\left(-\frac{c}{2}\left(z_0\sin\alpha+\frac{bv}{c}\right)^2\right)\right)$$

$$\left.+\mathrm{O}\left(\exp\left(-\frac{c}{2}\left(z_0\sin\alpha-\frac{bv}{c}\right)^2\right)\right)\right).$$

As $|v|$ is bounded by $y_0\sin\alpha$, the order of the error terms is $\mathrm{O}(\exp(-\frac{1}{2}c\sin^2\alpha(z_0-y_0b/c)^2)=\mathrm{O}(\exp(-\frac{1}{4}c(z_0-y_0b/c)^2\alpha^2))$, and the integral of $\mathrm{e}^{-u^2/2}$ extended to the real axis is equal to $\sqrt{2\pi}$. This finally gives for the integral in t:

$\sqrt{2\pi/c}\,(1+O(\exp(-\frac{1}{4}c(z_0-y_0b/c)^2)))$. We plug this value into the global integral $J(a,b,c)$ and we now have

$$J(a,b,c)=\sqrt{\frac{2\pi}{c}}\left(1+O\left(\exp\left(-\frac{c}{4}\left(z_0-y_0\frac{b}{c}\right)^2\alpha^2\right)\right)\right)$$
$$\times\int_{-y_0\sin\alpha}^{+y_0\sin\alpha}\exp(-(ac-b^2)\,v^2/2c)\,dv.$$

Now

$$\int_{-y_0\sin\alpha}^{+y_0\sin\alpha}e^{-kv^2/2}\,dv=(1/\sqrt{k})\left(\int_{-\infty}^{+\infty}e^{-v^2/2}\,dv+O(\exp(-(k/2)y_0^2\sin^2\alpha))\right)$$
$$=\sqrt{2\pi/k}(1+O(\exp(-(k/2)y_0^2\sin^2\alpha))).$$

We choose $k=(ac-b^2)/c$; this gives an approximation of $J(a,b,c)$ for $ac\neq b^2$:

$$J(a,b,c)=\frac{2\pi}{\sqrt{ac-b^2}}\left(1+O\left(\exp\left(-\frac{c}{2}\sin^2\alpha\left(z_0-y_0\frac{b}{c}\right)^2\right)\right)\right.$$
$$\left.+O\left(\exp\left(-\frac{k}{2}y_0^2\sin^2\alpha\right)\right)\right).$$

We now apply this formula to the integral $-J(h''_{y^2},h''_{yz},h''_{z^2})$: h''_{y^2} and h''_{z^2} are of order $\Theta(d)$ and h''_{yz} is of order $\Theta(d(x-1))=o(d)$; furthermore, the terms y_0 and z_0 are of order $\Theta(1)$. We can thus find a strictly positive constant C_2 such that the integral on $\gamma_{2,y}\times\gamma_{2,z}$ is equal to $-(2\pi)/\sqrt{ac-b^2}\,(1+O(e^{-C_2d\alpha^2}))$.
Summing up, we get

$$I'_1=-2\pi/\sqrt{h''_{y^2}h''_{z^2}-h''^2_{yz}}\,(1+O(e^{-C_2d\alpha^2})+O(d\alpha^3e^{-C_1d\alpha^2})+O(d\alpha^4e^{-C_0d\alpha^2})).$$

Putting all together, we obtain

$$I_1=\frac{e^{h(x,y(x),z(x))}}{2\pi\sqrt{h''_{y^2}h''_{z^2}-h''^2_{yz}}}(1+O(e^{-C_2d\alpha^2})+O(d\alpha^3e^{-C_1d\alpha^2})$$
$$+O(d\alpha^4e^{-C_0d\alpha^2})+O(d\alpha^3)).\tag{12}$$

Upper bound on I_2. I_2 is the integral of $e^{h(x,y,z)}$ on the set $[-\pi,+\pi]^2\backslash\mathscr{E}$:

$$I_2=\frac{1}{(2i\pi)^2}\int\int e^{h(x,y,z)}\,dy\,dz$$
$$=\frac{e^{h(x,y(x),z(x))}}{(2i\pi)^2}\int\int\exp(h(x,y(x)e^{i\theta},z(x)e^{i\tau})-h(x,y(x),z(x)))\,dy\,dz.$$

Define $k_x(\theta,\tau)=\lambda_R(y(x)e^{i\theta})+\lambda_R(xy(x)e^{i\theta})[\lambda_S(z(x)e^{i\tau})-1]$; we have

$$I_2=-\frac{y(x)z(x)e^{h(x,y(x),z(x))}}{4\pi^2}\int\int_{[-\pi,+\pi]^2\setminus\mathscr{E}}\left(\frac{k_x(\theta,\tau)}{k_x(0,0)}\right)^d e^{-i(r\theta+s\tau)}\,d\theta d\tau.$$

We want to find an upper bound on $|k_x(\theta,\tau)/k_x(0,0)|$ on the set $[-\pi,+\pi]^2\setminus\mathscr{E}$; this bound should not depend on x. Lemma 5.2 gives

$$|\lambda_R(y(x)e^{i\theta})|\leqslant\lambda_R(y(x))(1-C_1'\alpha^2),$$

with $C_1'>0$ and independent of x for $|\theta|\geqslant\alpha$. The term $|\lambda_R(xy(x)e^{i\theta})|^2$ satisfies a similar inequality, with a constant C_1'' independent of x and strictly positive:

$$|\lambda_R(x\,y(x)e^{i\theta})|\leqslant\lambda_R(x\,y(x))(1-C_1''\alpha^2).$$

- For $|\theta|\geqslant\alpha$, we have

  $$|k_x(\theta,\tau)|\leqslant|\lambda_R(y(x)e^{i\theta})|+|\lambda_R(x\,y(x)e^{i\theta})||\lambda_S(z(x)e^{i\tau})-1|.$$

  As the function λ_S has real positive coefficients and $\lambda_S(z)-1=\lambda_S(z)-\lambda_S(0)\geqslant 0$ for real positive z, $|\lambda_S(z(x)e^{i\tau})-1|$ is bounded by $\lambda_S(z(x))-1$ for all τ. This gives

  $$|k_x(\theta,\tau)|\leqslant k_x(0,0)-(C_1'\lambda_R(y(x))+C_1''\lambda_R(x\,y(x))(\lambda_S(z(x))-1))\alpha^2.$$

  We can assume that x is restricted to a compact subset around 1; the terms $\lambda_R(y(x))$ and $\lambda_R(x\,y(x))$ are then in a compact neighborhood of $\lambda_R(\rho_R)$, which is strictly positive: λ_R has positive coefficients and ρ_R is a positive real number. This means that the sum $C_1'\lambda_R(y(x))+C_1''\lambda_R(x\,y(x))(\lambda_S(z(x))-1)$ can be bounded away from 0 independently of x. Hence, there exists a constant $C_2'>0$ such that for any $|\theta|\geqslant\alpha$ and without restriction on τ;

  $$|k_x(\theta,\tau)|\leqslant k_x(0,0)(1-C_2'\alpha^2). \tag{13}$$

- To get an upper bound on $|\tau|\geqslant\alpha$, we have to extend Lemma 5.2 to the function $z\mapsto\lambda_S(z)-1$ (this requires that $\lambda_S(t)\neq 1+t$):

  $$|\lambda_S(z(x)e^{i\tau})-1|\leqslant(\lambda_S(z(x))-1)(1-C_1'''\alpha^2).$$

  We then show likewise that, for $|\tau|\geqslant\alpha$, and for any θ,

  $$|k_x(\theta,\tau)|\leqslant\lambda_R(y(x))+\lambda_R(x\,y(x))(\lambda_S(z(x))-1)(1-C_1'''\alpha^2),$$

  from which we deduce as above that

  $$|k_x(\theta,\tau)|\leqslant k_x(0,0)(1-C_2''\alpha^2) \tag{14}$$

  for a constant $C_2''>0$.

Let us define $C_3=\inf(C_2',C_2'')$; C_3 is strictly positive. The inequalities (13) and (14) show that, for any couple (θ,τ) of $[-\pi,+\pi]^2\setminus\mathscr{E}$, we have

$$|k_x(\theta,\tau)|\leqslant k_x(0,0)(1-C_3\alpha^2).$$

We get

$$|I_2| \leqslant \frac{y(x)z(x)e^{h(x,y(x),z(x))}}{4\pi^2} \int\int_{[-\pi,+\pi]^2 \setminus \mathscr{E}} (1-C_3\alpha^2)^d \, d\theta \, d\tau$$

$$\leqslant y(x)z(x)e^{h(x,y(x),z(x))}(1-C_3\alpha^2)^d.$$

Taking the main term of I_1 out of this approximation (cf. (12)), we get

$$|I_2| = \frac{e^{h(x,y(x),z(x))}}{\sqrt{h''_{y^2}h''_{z^2}-h''^2_{yz}}} O(de^{-C_3 d\alpha^2}). \tag{15}$$

Choice of α. We now choose α in such a way that the error terms in approximations (12) and (15) can be neglected. This is the case when α depends on d in such a way that the following conditions hold for $d \to +\infty$:

$$\frac{d\alpha^2}{\log d} \to +\infty, \quad d\alpha^3 \to 0.$$

For example, choose $\alpha = (\log d)/\sqrt{d}$, we get

$$\psi(x) = \frac{e^{h(x,y(x),z(x))}}{2\pi\sqrt{h''_{y^2}h''_{z^2}-h''^2_{yz}}}(1+o(1)). \tag{16}$$

5.3.3. Convergence of the Laplace transform towards $e^{t^2/2}$

We show here that the normalized Laplace transform $e^{t\mu/\sigma}\psi(e^{-t/\sigma})/\psi(1)$ converges towards $e^{t^2/2}$ when $d \to +\infty$ and for any t real positive. We have proved that the approximation (16) holds with $y(x)$ and $z(x)$, respectively, equal to the saddle points $\rho_R(1+v)$ and $\rho_S(1+u)$ and with values of h and its derivatives taken at point $(x, y(x), z(x))$. Define

$$\Xi(t) = \log\left(e^{t\mu/\sigma}\frac{\psi(e^{-t/\sigma})}{\psi(1)}\right) = t\mu/\sigma + \log\frac{\psi(e^{-t/\sigma})}{\psi(1)}.$$

Define also $\kappa(x) = (h''_{y^2}h''_{z^2} - h''^2_{yz})(x, y(x), z(x))$ and $\Delta h(x) = h(x, y(x), z(x)) - h(1, \rho_R, \rho_S)$. The approximation (16) gives for $x = e^{-t/\sigma}$:

$$\Xi(t) = t\mu/\sigma + \Delta h(e^{-t/\sigma}) - \frac{1}{2}\log\frac{\kappa(e^{-t/\sigma})}{\kappa(1)} + o(1).$$

We first study the term $\log \kappa(e^{-t/\sigma})/\kappa(1)$. In a neighborhood of 1, $\kappa(x) = \kappa(1) + O((x-1)\|\kappa'\|)$, κ is of order $\Theta(d)$, and $\kappa' = O(d)$. We have $\kappa(x)/\kappa(1) = 1 + O(x-1)$, and the logarithm reduces to an error term: $\log \kappa(x)/\kappa(1) = O(x-1)$. We next given an

expansion of $h(x,y,z)$. The derivatives are at point $(1,\rho_R,\rho_S)$; we recall that $h''_{yz}(1,\rho_R,\rho_S)=0$ and that, by choice of ρ_R and ρ_S, $h'_y(1,\rho_R,\rho_S)=h'_z(1,\rho_R,\rho_S)=0$:

$$h(x,y,z)=h(1,\rho_R,\rho_S)+(x-1)h'_x+\tfrac{1}{2}(x-1)^2h''_{x^2}+\tfrac{1}{2}(y-\rho_R)^2h''_{y^2}$$
$$+\tfrac{1}{2}(z-\rho_S)^2h''_{z^2}+(x-1)(y-\rho_R)h''_{xy}+(x-1)(z-\rho_S)h''_{xz}$$
$$+O(d(x-1)^3)+O(d(y-\rho_R)^3)+O(d(z-\rho_S)^3).$$

We substitute $y(x)=\rho_R(1+v)$ to y and $z(x)=\rho_S(1+u)$ to z; v and u are defined from $x-1$ by the formulae (10) and (11). For $C=h''_{x^2}-2\rho_R\alpha_1h''_{xy}-2\rho_S\alpha_2h''_{xz}+\rho_R^2\alpha_1^2h''_{y^2}+\rho_S^2\alpha_2^2h''_{z^2}$, we get

$$\Delta h(x)=h'_x(x-1)+\tfrac{1}{2}C(x-1)^2+O(d(x-1)^3).$$

These approximations of $\log(\kappa(x)/\kappa(1))$ and of $\Delta h(x)$ show that

$$\Xi(t)=t\mu/\sigma+h'_x(e^{-t/\sigma}-1)+\frac{C}{2}(e^{-t/\sigma}-1)^2+O\left(\frac{d}{\sigma^3}\right)+O\left(\frac{1}{\sigma}\right)+o(1),$$

which is

$$\Xi(t)=(\mu-h'_x)\frac{t}{\sigma}+(h'_x+C)\frac{t^2}{2\sigma^2}+O\left(\frac{d}{\sigma^3}\right)+O\left(\frac{1}{\sigma}\right)+o(1).$$

Define $\mu=h'_x$ and $\sigma^2=h'_x+C$. We easily check that

$$\mu=Ad\left(1-\frac{1}{\lambda_S(\rho_S)}\right),$$

$$\sigma^2=d\left((\rho_Rg'_R(\rho_R)+A^2)\frac{\lambda_S(\rho_S)-1}{\lambda_S^2(\rho_S)}-A^2\frac{\rho_S\lambda'^2_S(\rho_S)}{\lambda_R^4(\rho_S)g'_S(\rho_S)}\right).$$

In these formulae, ρ_R and ρ_S are the solutions of the equations $g_R(\rho_R)=(r+1)/d$ and $g_S(\rho_S)=(s+1)/d$; we can substitute the solutions of the limiting equations $g_R(y)=A$ and $g_S(z)=B$ for them, without changing the asymptotic order of the mean and variance: $\mu=d\mu_0$ and $\sigma^2=d\sigma_0^2$, with μ_0 and σ_0 strictly positive constants. The error terms are $O(1/\sqrt{d})$ and we get

$$\Xi(t)=\tfrac{1}{2}t^2+o(1).$$

5.4. Equijoin sizes: Theorem 3.6

The generating function associated with the equijoin has the general form

$$\Phi(x,y,z)=\left(\sum_{k,l\geq 0}a_kb_lx^{kl}y^kz^l\right)^d,$$

with coefficients $a_k=[t^k]\lambda_R(t)$ and $b_l=[t^l]\lambda_S(t)$. The coefficient of $y^r z^s$ in Φ can be computed with two applications of Cauchy's formula

$$\psi(x)=[y^r z^s]\Phi(x,y,z)=\frac{1}{(2i\pi)^2}\int\int e^{h(x,y,z)}\,dy\,dz,$$

with

$$h(x,y,z)=d\log\left(\sum_{k,l\geq 0} a_k b_l x^{kl} y^k z^l\right)-(r+1)\log y-(s+1)\log z$$

and where we integrate on circles around the origin. The function $\Phi(x,y,z)$ is not defined for all values of x, y and z. However, we shall work with x real in $]0,1]$; in this domain, $|\Phi(x,y,z)|\leqslant\sum_{k,l\geq 0} a_k b_l |y|^k |z|^l=\lambda_R(|y|)\lambda_S(|z|)$. Furthermore, the functions λ_R and λ_S are entire by the property $\mathscr{P}$, so the function Φ is well defined for all couples (y,z) and for x real in $]0,1]$.

5.4.1. Determination of the integration paths

The saddle points for $x=1$ are defined by the same equations as in the case of the semijoin:

$$g_R(y)=(r+1)/d,$$

$$g_S(z)=(s+1)/d.$$

We can solve these equations as soon as the usual limiting conditions (8) and (9) are satisfied. Let ρ_R and ρ_S be these solutions; they are the radii of the circles we choose for integration paths when $x=1$. We give here some information on the derivatives of h at point $(1,\rho_R,\rho_S)$: h'_y, h'_z and h''_{yz} are equal to 0 and the other derivatives are

$$h'_x(1,\rho_R,\rho_S)=g_R(\rho_R)g_S(\rho_S)d,$$

$$h''_{x^2}(1,\rho_R,\rho_S)=d(\rho_R g'_R(\rho_R)g_S^2(\rho_S)+\rho_S g'_S(\rho_S)g_R^2(\rho_R)$$
$$+\rho_R\rho_S g'_R(\rho_R)g'_S(\rho_S)-g_R(\rho_R)g_S(\rho_S)),$$

$$h''_{xy}(1,\rho_R,\rho_S)=d g'_R(\rho_R)g_S(\rho_S),$$

$$h''_{xz}(1,\rho_R,\rho_S)=d g'_S(\rho_S)g_R(\rho_R),$$

$$h''_{y^2}(1,\rho_R,\rho_S)=d\frac{g'_R(\rho_R)}{\rho_R},$$

$$h''_{z^2}(1,\rho_R,\rho_S)=d\frac{g'_S(\rho_S)}{\rho_S}.$$

We then choose the integration paths for $x=1+\varepsilon$. These paths are once more circles centered at the origin and whose radii are solutions of equations $h'_y(x,y,z)=0$ and

$h'_z(x, y, z)=0$. We define $y(x)=\rho_R(1+v)$ and $z(x)=\rho_S(1+u)$. Solving the approximate equations in ε, v and u gives

$$v=-g_S(\rho_S)\varepsilon+O(\varepsilon^2),$$

$$u=-g_R(\rho_R)\varepsilon+O(\varepsilon^2).$$

5.4.2. Approximation of the integral

The integration paths are the circles $y=y(x)e^{i\theta}$ and $z=z(x)e^{i\tau}$. The variable x is real near 1 and strictly smaller than 1. Let $\alpha\in]0,\pi/2[$; we choose $\alpha=(\log d)/\sqrt{d}$. Define $\mathscr{E}=]-\alpha,+\alpha[^2$. We show that the integral I_1 on $\mathscr{E}$ gives the important part and that the integral I_2 on $[-\pi,+\pi]^2\setminus\mathscr{E}$ can be neglected.

The evaluation of I_1 is similar to that given in Section 5.3.2 and we do not detail it here. We have $(h''_{y^2}h''_{z^2}-h''^2_{yz})(1,\rho_R,\rho_S)=d^2g'_R(\rho_R)g'_S(\rho_S)/(\rho_R\rho_S)$, hence,

$$(h''_{y^2}h''_{z^2}-h''^2_{yz})(x, y(x), z(x))=\frac{g'_R(\rho_R)g'_S(\rho_S)}{\rho_R\rho_S}d^2(1+O(x-1))\neq 0.$$

We obtain

$$I_1=\frac{e^{h(x,y(x),z(x))}}{\sqrt{h''_{y^2}h''_{z^2}-h''^2_{yz}}}(1+o(1)).$$

Using the function $k_x(\theta,\tau)=\sum_{k,l\geqslant 0}a_kb_lx^{kl}y(x)^kz(x)^le^{i(k\theta+l\tau)}$, we can write the integral I_2 as

$$I_2=\frac{e^{h(x,y(x),z(x))}}{4\pi^2}\iint_{[-\pi,+\pi]^2\setminus\mathscr{E}}\left|\frac{k_x(\theta,\tau)}{k_x(0,0)}\right|^d d\theta d\tau.$$

The main point in getting an upper bound on I_2 is, as before, to obtain a suitable upper bound on $|k_x(\theta,\tau)|$. But

$$k_x(\theta,\tau)=1+\{\lambda_R(y(x)e^{i\theta})-1\}+\{\lambda_S(z(x)e^{i\tau})-1\}$$

$$+\sum_{k,l\geqslant 1}a_kb_lx^{kl}y(x)^kz(x)^le^{i(k\theta+l\tau)}.$$

This gives

$$|k_x(\theta,\tau)|\leqslant 1+|\lambda_R(y(x)e^{i\theta})-1|+|\lambda_S(z(x)e^{i\tau})-1|+\sum_{k,l\geqslant 1}a_kb_lx^{kl}y(x)^kz(x)^l.$$

On $[-\pi,+\pi]^2\setminus\mathscr{E}$, at least one of the two following inequalities holds for a value C strictly positive and independent of d, r and s (see Lemma 5.2):[4]

$$|\lambda_R(y(x)e^{i\theta})-1|\leqslant(\lambda_R(y(x))-1)(1-C\alpha^2),$$

$$|\lambda_S(z(x)e^{i\tau})-1|\leqslant(\lambda_S(z(x))-1)(1-C\alpha^2).$$

[4] This is the place where the conditions $\lambda_R(t)\neq 1+t$, $\lambda_S(t)\neq 1+t$ are important.

We deduce that there exists a strictly positive constant C' such that, for all couples (θ, τ) of the set $[-\pi, +\pi]^2 \setminus \mathscr{E}$,

$$|k_x(\theta, \tau)| \leqslant k_x(0,0)(1 - C'\alpha^2).$$

This shows that $|I_2| = (e^{h(x,y(x),z(x))}/\sqrt{h''_{y^2}h''_{z^2} - h''^2_{yz}})\, O(de^{-C'd\alpha^2})$. We deduce from it an approximation of $\psi(x)$:

$$\psi(x) = \frac{e^{h(x,y(x),z(x))}}{\sqrt{h''_{y^2}h''_{z^2} - h''^2_{yz}}}(1 + o(1)).$$

5.4.3. Laplace transform and determination of moments

This part of the proof is very similar to that of the proof of Theorem 3.1 and we do not detail all the computations; the reader should have no difficulty in checking them.

The derivatives of h are $O(d)$ near point $(1, \rho_R, \rho_S)$ and $\kappa(x) = \sqrt{h''_{y^2}h''_{z^2} - h''^2_{yz}}$ satisfies $\log(\kappa(x)/\kappa(1)) = O(x-1)$. We then compute $\Delta h(x) = h(1, \rho_R, \rho_S) - h(x, y(x), z(x))$ and we get

$$\Delta h(x) = h'_x(x-1) + \left(h''_{x^2} - \frac{h''^2_{xy}}{h''_{y^2}} - \frac{h''^2_{xz}}{h''^2_{z^2}}\right)\frac{(x-1)^2}{2} + O(d(x-1)^3).$$

This proves the convergence of $e^{t\mu/\sigma}\psi(e^{-t/\sigma})/\psi(1)$ towards $e^{t^2/2}$, for any positive fixed t, and with the following values for μ and σ (we can substitute $g_R^{-1}(A)$ and $g_S^{-1}(B)$ for ρ_R and ρ_S):

$$\mu = dg_R(\rho_R)g_S(\rho_S) = ABd,$$

$$\sigma^2 = d\rho_R\rho_S g'_R(\rho_R)g'_S(\rho_S).$$

5.5. Equijoin of correlated relations: Theorem 3.10

The proof is very similar to that of Theorem 3.6 and we only indicate the points where it differs. The main difference is in the computation of the saddle points for $x = 1$. They are defined by

$$\frac{y\lambda'_R(y)(\lambda_S(z) - 1)}{1 + (\lambda_R(y) - 1)(\lambda_S(z) - 1)} = \frac{r+1}{d}, \tag{17}$$

$$\frac{z\lambda'_S(z)(\lambda_R(y) - 1)}{1 + (\lambda_R(y) - 1)(\lambda_S(z) - 1)} = \frac{s+1}{d}. \tag{18}$$

We cannot decompose this system in two equations, such that each of the equations has only one unknown variable. The fact that it was formerly possible to do so comes from the equality $\Phi(1, y, z) = \lambda_R(y)^d\lambda_S(z)^d$. However, the system (17), (18) still has real positive solutions under the usual assumptions on A and B.

Define ρ_R and ρ_S as the saddle points for $x=1$. Let

$$h(x,y,z)=d\log\left(1+\sum_{k,l\geqslant 1} a_k b_l x^{kl} y^k z^l\right)-(r+1)\log y-(s+1)\log z.$$

The saddle points for $x=1+\varepsilon$ are defined by

$$h''_{xy}\varepsilon+\rho_R h''_{y^2} v+\rho_S h''_{yz} t=O(\varepsilon^2),$$
$$h''_{xz}\varepsilon+\rho_R h''_{yz} v+\rho_S h''_{z^2} t=O(\varepsilon^2).$$

We compute a solution of the type $y=\rho_R(1+v)$ and $z=\rho_S(1+u)$, we get

$$\rho_R v=-\alpha\varepsilon+O(\varepsilon^2),$$
$$\rho_S u=-\beta\varepsilon+O(\varepsilon^2).$$

The coefficients α and β are functions of the derivatives of h or, equivalently, of the functions λ_R and λ_S, taken at point $(1,\rho_R,\rho_S)$:

$$\alpha=\frac{h''_{xy}h''_{z^2}-h''_{xz}h''_{yz}}{h''_{y^2}h''_{z^2}-h''^2_{yz}}(1,\rho_R,\rho_S),$$

$$\beta=\frac{h''_{xz}h''_{y^2}-h''_{xy}h''_{yz}}{h''_{y^2}h''_{z^2}-h''^2_{yz}}(1,\rho_R,\rho_S).$$

The rest of the proof is similar to that of Theorem 3.6.

This shows the convergence of the conditional probability distribution for the equijoin size, towards a normal distribution of asymptotic mean and variance $\mu=d\mu_0$ and $\sigma^2=d\sigma_0^2$, with suitable constants μ_0 and σ_0 for which we have (rather complicated) closed-form expressions. Once again, ρ_R and ρ_S can be changed to $g_R^{-1}(A)$ and $g_S^{-1}(B)$. The value μ_0 is

$$\mu_0\approx r\frac{\rho_S\lambda'_S(\rho_S)}{\lambda_S(\rho_S)-1}\approx s\frac{\rho_R\lambda'_R(\rho_R)}{\lambda_R(\rho_R)-1}.$$

The constant σ_0^2 can be written as a function of the derivatives of function $h(x,y,z)=d\log(1+\sum_{k,l\geqslant 1} a_k b_l x^{kl} y^k z^l)-(r+1)\log y-(s+1)\log z$, at point $(1,\rho_R,\rho_S)$:

$$\sigma^2=d\sigma_0^2=h'_x+h''_{x^2}+3\frac{h''^2_{xz}h''_{y^2}+h''^2_{xy}h''_{z^2}-2h''_{xy}h''_{yz}h''_{zx}}{h''_{y^2}h''_{z^2}-h''^2_{yz}}.$$

References

[1] H. Cartan, *Théorie élémentaire des fonctions analytiques d'une ou plusieurs variables complexes* (Hermann, Paris, 1961).

[2] S. Christodoulakis, On the estimation and use of selectivities in database performance evaluation, Tech. Report CS-89-24, Department of Computer Science, Univ. of Waterloo, Ontario, Canada, June 1989.

[3] M. Drmota, Asymptotic expansions and a multivariate Darboux method in enumeration problems, Tech. Report, Technical Univ. of Vienna, 1992.

[4] W. Feller, *An Introduction to Probability Theory and its Applications*, Vol. 2 (Wiley, New York, 1971).
[5] D. Gardy, *Bases de données, Allocations aléatoires: Quelques analyses de performances*, Thèse d'Etat, Université de Paris-Sud, June 1989.
[6] D. Gardy, Normal limiting distributions for projection and semijoin sizes, *SIAM J. Discrete Math.* **5** (1992) 219–248.
[7] D. Gardy and C. Puech, On the sizes of projections: a generating function approach, *Inform. Systems* **9** (1984) 231–235.
[8] D. Gardy and C. Puech, On the effect of join operations on relation sizes, *ACM Trans. Database Systems* **14** (1989) 574–603.
[9] I.P. Goulden and D.M. Jackson, *Combinatorial Enumeration* (Wiley, New York, 1983).
[10] M. Jarke and J. Koch, Query optimization in database systems, *ACM Comput. Surveys* **16** (1984) 111–152.
[11] N.L. Johnson and S. Kotz, *Urn Models and their Application* (Wiley, New York, 1977).
[12] V. Kolchin, B. Sevast'yanov and V. Chistyakov, *Random Allocations* (Wiley, New York, 1978).
[13] D. Maier, *The Theory of Relational Databases* (Computer Science Press, Rockville, MD, 1983).
[14] M.V. Mannino, P. Chu and T. Sager, Statistical profile estimation in database systems, *ACM Comput. Surveys* **20** (1988) 191–221.
[15] A.S. Rosenthal, Note on the expected size of a join, *SIGMOD Record* **11** (1991) 19–25.
[16] J.D. Ullman, *Principles of Database Systems* (Computer Science Press, Rockville, MD, 1980).