SIAM J. DISC. MATH.
Vol. 5, No. 2, pp. 219–248, May 1992

006

NORMAL LIMITING DISTRIBUTIONS FOR PROJECTION AND SEMIJOIN SIZES*

DANIÈLE GARDY†

Abstract. This paper presents classes of bivariate generating functions associated with the probability distributions of parameters on sets of points (sizes of derived relations in a relational data base) that correspond to asymptotically normal distributions. These results are extended to give some conditions under which the numbers $a_{n,k}$ defined by $\sum_{n,k} a_{n,k} x^k y^n = \phi(x,y)^d$ follow a Gaussian limiting distribution.

Key words. central limit theorem, generating function, urn models

AMS(MOS) subject classifications. 05A16, 60F05, 68P15, 68Q25

1. Introduction. The aim of this paper is to study some parameters that can be defined on sets of points obtained by random sampling without replacement from an initial domain Δ. We examine the probability distribution of these parameters, under the hypothesis that the size of the domain Δ and the sample size grow to infinity. More precisely, we assume that we know the probability distribution on Δ, and we want to show that, for a large class of these distributions, the probability distributions of the parameters that we consider are asymptotically normal.

Our tools for proving this convergence are the multivariate generating function of the parameter under study and the size(s) of the set(s) of points, and classical results on analytic functions and the Laplace transform. Bivariate generating functions for which the convergence toward a normal distribution holds are studied, for example, in [2], [4], [8]. Our approach differs from these works mostly in that the generating functions considered here depend on d, one of the parameters that grow to infinity. We consider *probability* or *counting* generating functions, which are themselves of the form "dth power of a function."

The plan of the paper is as follows: We present the parameters that we intend to study in §2. There we give several interpretations of these parameters, both probabilistic and related to data bases in computer science. We formally introduce our modelization and notations in §3. We next give our results in §§4 and 5 and prove them in §6.

2. Sets of points, relations, and sums of random variables.

2.1. Sets of points. We first define the set Δ of "legal" points. A point in a two-dimensional space is an ordered pair (x, y). We assume that each coordinate takes its value in a finite domain, denoted, respectively, by D_X and D_Y, on which a probability distribution is defined. Throughout the paper, d_X represents the size of the set D_X, and d_Y the size of the set D_Y. We also assume that the values of a point on its first and second coordinates are independent; i.e., the probability distribution on $\Delta = D_X \times D_Y$ is the product of the probability distributions on D_X and D_Y.

We consider a random subset R of Δ built in one of the following ways:

* Received by the editors December 2, 1988; accepted for publication (in revised form) April 30, 1991. This work was partially supported by the Programme de Recherches Concertées Mathématique-Informatique and by ESPRIT-II Basic Research Action No. 3075 (project ALCOM).

† LRI, Université de Paris-Sud, Centre National de la Recherche Scientifique U.A. 410, 91405 Orsay, France.

- In the first case, we obtain R by drawing n independent random points without replacement from Δ;
- We may also draw n independent random samples without replacement from D_X, then complete each pair by drawing independently the y-value from D_Y. In this case, each value of D_X appears at most once in a set R, but a given value of D_Y may be present in several pairs of R.

Of course, the symmetrical rule also exists: We can draw a sample without replacement from D_Y, then complete it by sampling from D_X. We define on R a first parameter $f(R)$: "number of distinct x-values" (or "number of distinct y-values"). The second parameter $g(R,S)$ in which we are interested is the size of the set of points obtained by drawing two independent sets R and S, then suppressing from R all the points (x,y) whose value x on the first coordinate does not appear in a pair of S. The sets R and S may take their values in the same sample space $D_X \times D_Y$ or in two different spaces $D_X \times D_Y$ and $D_X \times D_U$.

We want to investigate the relationship between the *size* of R (number of points in R) and $f(R)$, and between the sizes of R and S and $g(R,S)$, for different probability distributions on the domains D_X, D_Y, and D_U. More precisely, we are interested in the conditional probability distribution of $f(R)$ for a given size of R, and in that of $g(R,S)$ for given sizes of R and S. We study these distributions when the size of the domain D_X and the numbers of samplings (sizes of the sets of points R and S) grow to infinity, and we show that they become asymptotically normal in many cases.

2.2. Relational data bases and sizes of relations. Those who are familiar with that part of computer science that deals with relational data bases may have noticed that the sets of points R and S defined in §2.1 are instances of relations of a particular type. The coordinates X and Y, or X and U, are the so-called *attributes* of relations R and S, and the points are the tuples of the relations. The parameters $f(R)$ and $g(R,S)$ are, respectively, the sizes of the *projection* of relation R on attribute X (or attribute Y) and of the *semijoin* of relations R and S on attribute X. These sizes are important parameters in query optimization, which aims at minimizing the cost of executing a query on the data base. We refer to [18], [20] for general texts on relational data base theory, to [15], [19] for surveys on query optimization and on the evaluation of relation sizes, and to [9], [11], [12] for a complete presentation of the problem of relation sizes and its modelization in terms of generating functions. We mention in [12] that the probability distributions of the sizes of relations obtained by a projection or a semijoin were (empirically) found to follow asymptotically normal distributions. Here we make precise the conditions under which this convergence holds and give the mathematical proofs. We also prove that complete knowledge of the probability distributions on all attributes is not necessary to characterize the asymptotic distribution of the derived size, and that it often suffices to know the distribution on the domain of the attribute on which the projection or the semijoin takes place.

The classical operations defined on relational data bases are the set operations (intersection, union, and symmetrical difference), the projection, and several types of join, mostly the equijoin and the semijoin [18], [20]. We restrict this paper to the *projection* and *semijoin*. The *intersection* is related to a special case of semijoin, and the sizes of the *union* and *difference* are very easily computed from the sizes of the intersection and of the initial relations. We do not consider the *equijoin* in this paper. One justification is that query optimizers often use a sequence of semijoins to reduce data before computing a "full" equijoin, and that an important part of

the cost of the operation comes from the semijoin part. We must also admit that the generating functions associated with the equijoins are less easy to study than the functions associated with the semijoins, and we defer them to a forthcoming paper [10].

We assume that the relations we consider have two (sets of) attribute(s): X and Y or U. Throughout the paper, X denotes the join or projection attribute. We restrict ourselves to the following three schemes of relations:

- In the case of a *free* relation, there is total independence between the values taken by the different tuples. This is the first case of §2.1;
- We may also consider relations where attribute X is a *key*, i.e., in a given instance of the relation the x-value of a tuple uniquely determines its y-value. This is the second case of §2.1;
- Finally, we consider the symmetrical case, where attribute Y is key of relation R.

Of course, there are many more possible schemes of relations. We give in [9], [12] generating functions for several of them.

2.3. Sums of random variables. It can be recognized that the parameters $f(R)$ and $g(R,S)$ defined in §2.1 are instances of a common problem: We study the limiting distribution of a sum of identically distributed *dependent* random variables when the number of variables grows to infinity.

Given two sets R and S built as described in §2.1, we define two random variables for each i in D_X: v_i and w_i are, respectively, the number of points of R or S whose value on the first coordinate is i. These variables take their values in $\{0 \cdots d_Y\}$, and the case where i does not appear in a pair of, say, R corresponds to $v_i = 0$. The sizes of R and S can be expressed as

$$\sum_{1 \le i \le d_X} v_i \quad \text{and} \quad \sum_{1 \le i \le d_X} w_i.$$

The size of the projection of R on the first coordinate is

$$\sum_{1 \le i \le d_X} u_i, \quad \text{with } u_i = 1_{v_i > 0}.$$

The size of the semijoin of relations R and S can also be written as

$$\sum_{1 \le i \le d_X} u_i', \quad \text{with } u_i' = v_i \cdot 1_{w_i > 0}.$$

If we assume a uniform probability distribution on the domain D_X, then the random variables $u_i, 1 \le i \le d_X$ (or the u_i') follow an identical distribution. Our problem, then, is to study the sum of the u_i, or the sum of the u_i', under the conditions that the sums of the v_i and w_i are known and when the total number d_X of variables grows to infinity.

In the case of *independent* random variables u_i, the central limit theorem, or some extensions of it when the variables are not uniformly distributed (see, for example, [13]), allows us to prove that the distribution of the sum $\sum_{1 \le i \le n} u_i$ is asymptotically normal for large n. We see here that, although the random variables are no longer independent, the correlation between them is weak enough that the limiting distribution is still Gaussian.

3. Models and notations.

3.1. Probability distributions on attribute domains. We consider two classes of distributions on a finite domain D of size d and denote by $p_{i,d}$ the probability that the ith element of the domain is selected when choosing at random an element of D. Hence we have that $\sum_i p_{i,d} = 1$. The subscript emphazises the fact that the probability distribution depends on the number d of elements in the domain. Without loss of generality, we can assume that the $p_{i,d}$ are decreasing when i grows, for fixed d. The two classes are defined as follows:

(Z) $\sum_{1\le i\le d} {p_{i,d}}^2 \to 0$ 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.

Class (Z) is named after the *Zipf* distribution: $p_{i,d}$ proportional to $1/i^C$ for $1 \le i \le d$ and fixed d. The uniform probability distribution, the so-called "80% – 20%" distributions and Zipf distributions for $0 < C \le 1$ are members of this class. A probability distribution on a domain of fixed size d_0 ($p_{i,d} = p_{i,d_0}$ for $i \le d_0$ and $p_{i,d} = 0$ for $i > d_0$), Zipf distributions for $C > 1$, and geometric distributions belong to class (G). Intuitively, distributions of class (Z) are not too far from the equiprobable case, and distributions of class (G) are those for which the probability of the "diagonal" $\{(i,i)\}$ has a nonnull limit.

Distributions of classes (Z) and (G) share the uniform convergence, for bounded t, of the generating function $\lambda(t) = \prod_{1\le i\le d}(1 + p_{i,d}t)$ associated with the probabilities of the sets of distinct items, toward a function $\varphi(t)$.[1] Probability distributions of class (Z) are simply characterized by $\varphi(t) = e^t$. Anticipating the results presented below, we can see that the distributions on the attributes that do not participate in the projection or in the join (attributes Y and U) matter only as long as they belong either to class (Z) or to class (G). In particular, all distributions of class (Z) give distributions for the projection or semijoin size that converge asymptotically to the same normal distribution, uniquely characterized by its moments.

3.2. Probability distributions on relations. We recall the independence assumptions of §2.1, translated in terms of the following relations:

(i) The two coordinates of a tuple (point) are independent;

(ii) The tuples (points) of a given relation (set) are independent, as far as this is compatible with the constraints on the relation (free relation or relation with a key);

(iii) When we consider two relations R and S, these two relations are independent.

Condition (i) ensures that the probability distribution on a domain $\Delta = D_X \times D_Y$ is the product of the probability distributions on domains D_X and D_Y, and condition (iii) merely states that the probability distribution of a couple (R, S) is the product of the probabilities of R and S. Condition (ii) was detailed in §2.1 and deserves further explanation.

The underlying idea is that the probability distribution on a relation R is proportional to the probability of each of its points: $\mathrm{Prob}(R) = k \cdot \prod_{t\in R} \mathrm{Prob}(t)$. The constant k is independent of R and is chosen to obtain a probability distribution on relations; it varies according to the rule for building R, which may restrict the set of

[1] The generating function that gives the probabilities of the finite sets of elements is actually $\lambda_0(t) = \prod_{1\le i\le d} \big((1 + p_{i,d}t)/(1 + p_{i,d})\big)$. It differs from the function $\lambda(t)$ that we use in the paper by a constant multiplicative factor $\prod_{1\le i\le d}(1 + p_{i,d})$, which disappears when we study conditional distributions; see §3.3.

admissible relation instances. For example, assume that the probability distribution on attribute X is given by $\{q_j, 1 \leq j \leq d_X\}$ and that the distribution on attribute Y is given by $\{p_i, 1 \leq i \leq d_Y\}$; in the case of a free relation (first case of §2.1), we have that $k = 1/\prod_{i,j}(1 + p_i q_j)$; in the case of a relation with key X (second case of §2.1), we have that $k = 1/\prod_j(1 + q_j)$ [12].

3.3. Limiting distributions. We recall that we want to investigate the limiting distribution of the size of a relation obtained either by the projection of a relation of known size r or by the join of two relations of known sizes r and s. Thus the problems presented in this paper can be cast into a common frame: given a doubly indexed sequence of real positive numbers $(a_{l,r})$,[2] our goal is to study the limiting distribution of the normalized sequence $(b_{l,r} = a_{l,r}/(\sum_l a_{l,r}))$ when r goes to infinity. We assume that we know the function $\Phi(x, y) = \sum_{l,r} a_{l,r} x^l y^r$. The problem can be reformulated using the probability distribution defined by the generating function $f(x) = [y^r]\Phi(x, y)/[y^r]\Phi(1, y)$: This is the conditional probability distribution of the parameter "marked" by x in Φ, knowing that the parameter "marked" by y in Φ (usually the size of some structure) has value r. We study the limit of this conditional distribution when r and d go to infinity.

For example, we define $a_{l,r}$ as the number of relations of size r whose projection is of size l, and we want to estimate the size of the projection of a relation of known size r; d is the size of the domain on which we project the relation. We see in §4 that the generating function that appears in the study of the projection size has the general form $\Phi(x, y) = (1 - x + x\lambda(y))^d$.

In this form, it is obvious that, at least for uniform distributions on the underlying domains, it does not matter if Φ is a probability or counting generating function in the variable y: This corresponds to an extra factor in the term $[y^r]\Phi(x, y)$, which cancels in $f(x)$. In our example, the generating function for the projection sizes might be a probability generating function with respect to x, and a counting generating function with respect to y. For the same reason, we may indifferently use an ordinary or exponential function in y, according to the underlying structure (this holds even if the distribution on the attribute domains are not uniform). Finally, we use probability generating functions Φ either for joint probabilities or for conditional probabilities (we assume that we know the size of the parameter marked by y) as the need arises: The generating function for the conditional probability satisfies $[y^r]\Phi(1, y) = 1$, which gives $f(x) = [y^r]\Phi(x, y)$.

We give our theorems in the case where *the probability distribution on the domain D_X of the projection or join attribute* X *is uniform,* and its size d_X [3] is related in a simple way to the sizes of the relevant relations. For example, in Corollary 1 the size d_X of D_X is of the order of the size r of relation R. We also assume that, when the sizes d_Y and d_U of the domains D_Y and D_U grow to infinity, they do so without relation to d_X. However, the exact relation between these parameters is not strict: The proofs can be adapted in many cases to show the convergence toward normal distributions with suitably modified moments.

3.4. Analytic functions with positive coefficients. For easy reference, we introduce here a property relative to an analytic function that we need to prove

[2] The numbers $a_{l,r}$ actually depend on a third parameter d in such a way that the function $\Phi(x, y) = \sum_{l,r} a_{l,r} x^l y^r$ is of the form $\phi(x, y)^d$. See §§4 and 5.

[3] This is the parameter d such that $\Phi = \phi^d$.

our results and which is satisfied in all the cases studied in this paper; we call it Property $\mathcal{P}$, shown below.

PROPERTY $\mathcal{P}$. *A function, say $\lambda(y)$, is entire and not affine, with positive coefficients, such that $\lambda(0) = 1$, and such that there exists no entire function Λ and integer $m \geq 2$ with $\lambda(y) = \Lambda(y^m)$.*

The last part of Property $\mathcal{P}$, $\lambda(y) \neq \Lambda(y^m)$, is introduced for technical reasons, but is in no way a restriction: If $\lambda(y) = \Lambda(y^m)$, we just change the variable y into y^m for the greatest such m, and the function Λ satisfies Property $\mathcal{P}$. It can be reformulated as in [4]: The greatest common divisor of the $\{r : [y^r]\lambda \neq 0\}$ is 1. Likewise, the important condition on $\lambda(0)$ is simply $\lambda(0) \neq 0$; requiring that $\lambda(0) = 1$ merely simplifies some computations.

In Theorem 1 of §4 and in Theorems 3 and 4 of §5, we use an auxiliary function $g(y)$, obtained from the function $\lambda(y)$ by $g(y) = y\lambda'(y)/\lambda(y)$.

LEMMA A. *Let Property $\mathcal{P}$ be satisfied and define $g(y) = y\lambda'(y)/\lambda(y)$. Then g is increasing on the interval $[0, +\infty[$.*

Proof of Lemma A. As function λ is entire with positive coefficients and $\lambda(0) = 1$, λ has no zeros on $[0, +\infty[$, and the function g is well defined on this interval. Let us define the function

$$D(y) = \lambda(y)(\lambda'(y) + y\lambda''(y)) - y\lambda'^2(y).$$

We have that $g'(y) = D(y)/\lambda^2(y)$. The definition of D in terms of λ and its derivatives can be used to get an expansion of D as a series with positive terms. This shows that g' is positive on the interval $[0, +\infty[$ and that g is increasing on this interval. □

When the function λ satisfies Property $\mathcal{P}$, then, by Lemma A, the function g is increasing; hence $g(y)$ either has a finite limit or tends to infinity when $y \to +\infty$. Henceforth, we use the expression $\lim_{y\to+\infty} g(y)$ either for a finite or an infinite limit; in the last case, the condition that the limit is greater than some positive number A is trivially satisfied.

4. Asymptotic distributions for projection sizes. Given a relation R with two attributes X and Y, we want to study the size of the *projection of* R *on attribute* X. We recall that this projection is computed by suppressing the attribute Y, then eliminating the redundant values of attribute X: We just keep one instance of each x-value that appears in the initial relation. We assume that the domains D_X and D_Y, where X and Y take their values, are of finite sizes d_X and d_Y and that the relation has r elements, where r is of the order of d_X. We are interested in the probability distribution of the size of the projection of R, conditioned by the initial size r of R, when the parameters r, d_X, and d_Y grow to infinity.

To be consistent with the schemes of relations defined in §2.1, we study a relation with a key on the attribute Y eliminated by the projection and a relation without a key. The case of the projection of a relation R with a key on attribute X is without any difficulty: Each pair of R has for X-component a distinct value; as a consequence, the projection on attribute X is composed of all the values x that appear as the first coordinate of a pair (x, y), counted once, and has exactly the same size as the initial relation R.

4.1. R has Y as key. Define $p(l/r)$ as the conditional probability that the projection of R on attribute X is of size l when the size of R is itself equal to r, for a uniform probability distribution on attribute X and a general probability distribution on attribute Y given by $\{p_{i,d_Y}, 1 \leq i \leq d_Y\}$. To study the distribution

of the projection size, it is convenient to use the following generating function, exponential in y: $\Phi(x,y) = \sum_{l,r} p(l/r)x^l y^r/r!$. As an intermediate step, we use the auxiliary bivariate generating function $\Psi(x,y) = \sum_{l,r} p(l,r)x^l y^r$. In both functions Φ and Ψ, the variable x marks the size of the projection on attribute X, and the variable y the size of the initial relation; $p(l,r)$ is the joint probability that relation R is of size r and its projection on attribute X is of size l; it is related to $p(l/r)$ by $p(l/r) = p(l,r)/(\sum_k p(k,r)) = [x^l y^r]\Psi(x,y)/[y^r]\Psi(1,y)$. We have [9], [11] that

$$\Psi(x,y) = \sum_{k=0}^{d_X} \binom{d_X}{k} \lambda_0(ky/d_X)\, x^k\, (1-x)^{d_X-k}.$$

In this formula, $\lambda_0(t) = \prod_{1\le i\le d_Y} ((1+p_{i,d_Y}t)/(1+p_{i,d_Y}))$ is the generating function describing all sets of y-values, with their associated probability. By extracting the coefficient of y^r in Ψ and computing the (exponential in y) generating function of the conditional probabilities $\Phi(x,y) = \sum_{l,r} p(l/r)x^l y^r/r!$, we get [11] that

$$\Phi(x,y) = (1 + x(e^{y/d_X} - 1))^{d_X}. \tag{1}$$

Let us mention that there exists a closed-form expression for the conditional probabilities: $p(l/r) = l!\binom{d_X}{l} d_X^{-r} S(r,l)$, with $S(r,l)$ a Stirling number of the second kind.[4]

Equation (1) shows that the evaluation of the projection size for a relation, with a key on the attribute Y suppressed by the projection, is equivalent to the classical *occupancy problem in urn models* [16]. This problem can be summarized as follows: Given d urns and r balls, the balls are thrown independently and at random into the urns, and we study the number of empty urns, or, equivalently, the number of urns containing at least one ball. The appropriate generating function in this case is a counting generating function, exponential in the number of balls (marked by the variable y) and ordinary in the number of urns with at least one ball (marked by x). Let us denote by $N_{l,r}$ the number of ways of throwing r balls into l urns, with each urn containing at least one ball; then [16] it follows that

$$\sum_{l,r} N_{l,r} x^l y^r/r! = (1 - x + xe^y)^d.$$

It is obvious from the expression of $\Phi(x,y)$ in (1) (but not from that of Ψ) that the probability distribution on attribute Y does not matter. Moreover, a whole spectrum of limiting results is known for urn models (see [16], [17] for surveys) and can be directly applied to the projection of a relation with a key.

4.2. R is a free relation. The bivariate generating function $\Phi(x,y)$ of the joint probabilities $p(l,r)$, where x marks the size of the projection on attribute X and y the size of the initial relation, is [9], [11]

$$\Phi(x,y) = \sum_{l,r} p(l,r)x^l y^r = (1 - x + x\lambda(y))^{d_X},$$

[4] As the notation for Stirling numbers is not standardized, we use here the notation of Comtet [5].

R	$\Phi(x,y)$	Asymptotic result
$X \dagger Y$	$(1-x+x\lambda(y))^{d_X}$	§4.2, Cor. 1
$Y \rightarrow X$	$(1-x+xe^{y/d_X})^{d_X}$	[16], [17] or §4.2, Thm. 1

FIG. 1. *Generating function for the size of the projection $\pi_X(R)$ of R on attribute X : $\pi_X(R) = \{x | \exists y : (x,y) \in R\}$.*

with

$$\lambda(y) = \prod_{1 \leq i \leq d_Y} (1 + p_{i,d_Y}\ y).^5$$

In this formula, as in §4.1, p_{i,d_Y} denotes the probability of the ith value of domain D_Y, which depends on the type of distribution and on the size of the domain.

Figure 1 sums up the generating function and the asymptotic results, either previously known or proved in this paper, for the two types of relations: a free relation and a relation with a key. Here and in Fig. 2 in §5.1, "X † Y" means that neither attribute X nor attribute Y is key of R (free relation), $Y \rightarrow X$ means that the attribute Y is key of R, and $X \rightarrow Y$ (in Fig. 2) means that X is key of R.

We first give a general theorem (Theorem 1) pertaining to functions that have the general form $(1-x+x\lambda(y))^d$. We then deduce from it a corollary dealing with the case of a free relation. Theorem 1 can also be used to get the classical result on urn models, or, equivalently, the result pertaining to a relation where attribute Y is key: This is simply the case where the function $\lambda(y)$ is equal to e^y or to e^{y/d_X}.

THEOREM 1. *Let Property $\mathcal{P}$ be satisfied. Define $\Phi(x,y) = (1-x+x\lambda(y))^d$. Let $d, r \rightarrow +\infty$ in such a way that $r = Ad + o(d)$ for some positive constant A, and that $g(y) = y\lambda'(y)/\lambda(y)$ satisfies $\lim_{y \rightarrow +\infty} g(y) > A$. Then the probability distribution defined by the generating function $f(x) = [y^r]\Phi(x,y)/[y^r]\Phi(1,y)$ is asymptotically Gaussian when $d \rightarrow +\infty$. The asymptotic values of the mean and variance are defined in terms of the unique real positive solution ρ of the equation $g(y) = A$ as follows :*

$$\mu = d\left(1 - \frac{1}{\lambda(\rho)}\right); \qquad \sigma^2 = d\left(\frac{1}{\lambda(\rho)} - \frac{1}{\lambda^2(\rho)} - \frac{\rho\lambda'^2(\rho)}{g'(\rho)\lambda^4(\rho)}\right).$$

COROLLARY 1. *Let $R[X,Y]$ be a free relation with a uniform probability distribution on the domain of attribute X Then the probability distribution of the size of the projection of R on attribute X, conditioned by the size $r = Ad_X + o(d_X)$ of relation R*

[5] Actually, $\Phi(x,y)$ is obtained from the function

$$\lambda_0(t) = \lambda(t)/\lambda(1) = \prod_i \left((1 + p_{i,d_Y} y)/(1 + p_{i,d_Y})\right)$$

by marking the tuples of R and their projection on X; this gives

$$\Phi(x,y) = \left((1-x+x\prod_i (1+p_{i,d_Y} y))/\prod_i (1+p_{i,d_Y})\right)^{d_X} = \frac{(1-x+x\lambda(y))^{d_X}}{\lambda(1)^{d_X}}.$$

As we are interested in $f(x) = [y^r]\Phi(x,y)/[y^r]\Phi(1,y)$, the multiplicative factor $\lambda(1)^{d_X}$ cancels in $f(x)$, and we can use the simpler expression given in the text.

with A a positive constant is asymptotically normal when $d_X \to +\infty$. The asymptotic mean and variance are given by $\mu = \mu_0\, d_X$ and $\sigma^2 = \sigma_0^2 d_X$ where μ_0 and σ_0^2 are constants that depend on the probability distribution on attribute Y. We now assume that $d_Y \to +\infty$ and is independent of d_X.

If the distribution on D_Y satisfies hypothesis (Z), *then $\mu_0 = 1 - e^{-A}$ and $\sigma_0^2 = (e^A - 1 - A)/e^{2A}$. If the distribution on D_Y satisfies hypothesis* (G), *let $\varphi(t) = \prod_{i\geq 1}(1 + p_i t)$, where the $\{p_i\}$ define the limiting distribution on attribute Y, and $g(t) = t\,\varphi'(t)/\varphi(t)$. Let ρ be the unique real positive solution of the equation $g(t) = A$. The constants μ_0 and σ_0^2 are*

$$\mu_0 = 1 - 1/\varphi(\rho); \qquad \sigma_0^2 = \frac{\mu_0}{\varphi(\rho)} - \frac{\rho\varphi'^2(\rho)}{g'(\rho)\varphi^4(\rho)}.$$

Moreover, μ_0 satisfies $1 - e^{-A} \leq \mu_0 \leq 1$.

The proof of Theorem 1 and the derivation of Corollary 1 are postponed until §6.2. As an application of Corollary 1, we deduce that the exact probability distribution on attribute Y has no influence on the limiting distribution as long as it stays in class (Z) and $d_Y \to +\infty$.

Relation to some urn models. When the probability distribution on the domain of attribute Y belongs to class (Z), i.e., when the function

$$\lambda(y) = \prod_{1\leq i\leq d_Y} (1 + p_{i,d_Y} y)$$

has for limit e^y for any fixed y and for $d_Y \to +\infty$, the generating function $\Phi(x, y) = (1 - x + x\prod_{1\leq i\leq d_Y}(1 + p_{i,d_Y} y))^{d_X}$ converges pointwise toward the function $(1 + x(e^y - 1))^{d_X}$ when d_Y grows to infinity and d_X is constant. This function is the generating function $\sum_{i,j} N_{i,j} x^i y^j / j!$ of the number i of urns containing at least one ball when we throw j balls independently in d_X urns, and it has already appeared in the study of a relation with a key (see §4.1). This can be explained intuitively as follows: For large d_Y, the probability $\sum_i p_{i,d_Y}^2$ that we twice draw the same point in successive trials with replacements is close to zero. Hence we may assume that the successive trials that give the points of the relation are "asymptotically" independent, and we get the classical urn model.

However, when the probability distribution on attribute Y belongs to class (G), the successive trials giving the points of the relation are *not* independent: The probability of twice drawing the same point in random sampling with replacement is definitely not null! (Asymptotically, it is close to $\sum_{i\geq 1} p_i^2 > 0$.) This is reflected in the limiting generating function $(1 + x(\varphi(y) - 1))^{d_X}$, which we obtain by letting d_Y grow to infinity and by keeping d_X constant.

Alternatively, the size of the projection on attribute X can be related to the number of nonempty urns when we throw the balls in *complexes*. Again, we refer to [17] for asymptotic results when complexes are of fixed size. In our approach, a complex is the number of points (x_0, y) in a given instance of a relation for a fixed value x_0 of attribute X, and its size is a random variable taking its values in $\{0 \cdots d_Y\}$. Ammann [1] studies such a case when the size of a complex is bounded and for various conditions on the numbers r of balls and d of urns: If r is of order $\sqrt{d}$, the number of empty urns asymptotically follows a compound Poisson distribution; for larger r, but still with $r = o(d)$, the asymptotic distribution becomes Gaussian. In our framework, this means that the size of domain D_Y is fixed and that the order of the size of the relation is either $\sqrt{d_X}$ or $o(d_X)$.

5. Asymptotic distributions for semijoin sizes. In this section, we consider two initial relations R and S and their semijoin on a common attribute X. The values taken by the two relations are assumed to be independent of each other. We consider that R and S are each built on two attributes, respectively, $R[X, Y]$ and $S[X, U]$. The *semijoin* of R and S on attribute X is a subset of the relation R; it is computed by keeping in relation R only those tuples whose value on X appears in the X-column of relation S. This operation is not symmetrical: The semijoin of R and S is *not* equal to the semijoin of S and R. We recall that we assume a uniform probability distribution on the join attribute X.

5.1. Generating functions. We use the following notation: $p(t, r, s)$ is the joint probability that the relations R and S have respective sizes r and s and that their semijoin is of size t; $p(t, s/r)$ is the joint probability that the relation S is of size s, and that the semijoin of R and S is of size t, conditioned by the fact that the relation R is of size r; and so forth. These probabilities are trivially related to one another; for example, $p(t/r, s) = p(t, r, s)/(\sum_i p(i, r, s))$. We define two functions λ_R and λ_S as in §3.1. For example, λ_R is a generating function associated with the sets of elements of R whose first value is fixed. If R is a free relation and if the probability distribution on attribute Y is given by $\{p_{i,d_Y}\}$, then $\lambda_R(y) = \prod_{1 \le i \le d_Y} (1 + p_{i,d_Y} y)$. When X is the key of R, then $\lambda_R(y) = 1 + y$. Once again, we do not consider the probability generating function, which only differs from λ_R by a constant factor. Function λ_S describes in a similar way the legal sets of points in S.

Our aim is to study the conditional probability distribution $p(t/r, s)$, which gives the probability that the join has size t, knowing that the initial relations are, respectively, of sizes r and s. As in §4.1, we often use as an intermediary step the generating function of another, related distribution. We use whatever probability distribution has a generating function of a kind convenient for asymptotic study, namely $\phi(x, y, z)^d$. The rule of thumb is that, if an attribute Y or U is the key of the relation in which it appears (R or S), we should use a probability distribution conditioned by the size of this relation; moreover, the generating function should be exponential in the variable "marking" the relation. This is formalized in Theorem 2, below.

THEOREM 2. *Let R and S be two independent relations. The generating function $\Phi(x, y, z)$ of the sizes of relations R and S, and of the semijoin of R and S, is given by the table of Fig. 2, with the conventions that "X † Y" or "X † U" corresponds to a free relation, $Y \to X$ or $U \to X$ to a relation with key Y or U as applies, and $X \to Y$ or $X \to U$ means that X is key of R or S, and with the following definition of Φ:*

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

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

- *If attribute Y is key of relation R, and relation S is either free or with key X,*

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

- *If relation R is free or has key X, and attribute U is key of relation S,*

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

R	S	$\Phi(x,y,z)$	Asymptotic result
$X \dagger Y$	$X \dagger U$	$(\lambda_R(y) + \lambda_R(xy)[\lambda_S(z) - 1])^{d_X}$	[10]
$X \dagger Y$	$X \rightarrow U$	$(\lambda_R(y) + z\lambda_R(xy))^{d_X}$	§5.4, Cor. 4
$X \dagger Y$	$U \rightarrow X$	$(\lambda_R(y) + \lambda_R(xy)[e^z - 1])^{d_X}$	[10]
$X \rightarrow Y$	$X \dagger U$	$(1 + y + (1 + xy)[\lambda_S(z) - 1])^{d_X}$	§5.3, Cor. 2
$X \rightarrow Y$	$X \rightarrow U$	$(1 + y + z + xyz)^{d_X}$	§5.4, Thm. 5
$X \rightarrow Y$	$U \rightarrow X$	$(1 + y + (1 + xy)[e^z - 1])^{d_X}$	§5.3, Cor. 3
$Y \rightarrow X$	$X \dagger U$	$(e^y + e^{xy}[\lambda_S(z) - 1])^{d_X}$	[10]
$Y \rightarrow X$	$X \rightarrow U$	$(e^y + ze^{xy})^{d_X}$	§5.4, Cor. 5
$Y \rightarrow X$	$U \rightarrow X$	$(e^y + e^{xy}[e^z - 1])^{d_X}$	[10]

FIG. 2. *Generating function for the size of the relations R, S, and their semijoin on attribute* X: $\{(x,y)|(x,y) \in R; \exists u : (x,u) \in S\}$.

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

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

Proof of Theorem 2. The ordinary counting generating functions for the cases when none of the attributes Y and U is key can be found in [12]. The computation of the *joint probability* generating functions when neither attribute Y nor attribute U is key of its relation is straightforward, and we do not detail it. We give below the computation of $\Phi(x,y,z)$ when attribute Y is key of relation R and relation S is free. The cases where relation S has for key either X or U can be dealt with in a similar way.

We first assume that the probability distributions on attributes Y and U are uniform. It is simpler in this case, and it has no effect on function Φ, to use the *counting* generating function of the sets of elements on attribute X, that is, to take $\lambda_S(y) = (1+z)^{d_Z}$ instead of $(1 + z/d_Z)^{d_Z}$. Let us denote by $N(t,r,s)$ the number of couples of relations (R,S) with given sizes r and s, whose semijoin has size t, and by $N(r)$ the number of relations R of size r. The ordinary counting generating function $\Psi(x,y,z) = \sum_{r,s,t} N(t,r,s)\; x^t y^r z^s$ is [12]

$$\Psi(x,y,z) = \sum_k \binom{d_X}{k} (1 + d_X y + ky(x-1))^{d_Y} (\lambda_S(z) - 1)^k.$$

The conditional probability $p(t,s/r)$ is equal to $N(t,r,s)/N(r)$. We want to compute $\Phi(x,y,z) = \sum_{r,s,t} p(t,s/r)\; x^t \cdot y^r/r! \cdot z^s = \sum_{t,r,s} N(t,r,s)/N(r) \cdot x^t \cdot y^r/r! \cdot z^s$. Substituting the value $d_X^r \binom{d_Y}{r}$ for $N(r)$ in the expression of $p(t,s/r)$ gives

$$p(t,s/r) = \frac{[x^t y^r z^s]\Psi(x,y,z)}{d_X^r \binom{d_X}{r}} = \sum_k \binom{d_X}{k} d_X^{-r} [x^t](d_X - k + kx)^r [z^s](\lambda_S(z) - 1)^k.$$

Substituting this value into the definition of Φ, we get that

$$\Phi(x,y,z) = \sum_{k,r,s,t} \binom{d_X}{k} [x^t]\{(d_X - k + kx)^r\} x^t \cdot y^r/(d_X^r r!) \cdot [z^s]\{(\lambda_S(z) - 1)^k\} \cdot z^s$$

$$= e^y \sum_k \binom{d_X}{k} e^{k(x-1)y/d_X} \cdot (\lambda_S(z)-1)^k = (e^{y/d_X} + e^{xy/d_X} \cdot [\lambda_S(z)-1])^{d_X}.$$

The computation when at least one of the probability distributions on attributes Y or U is not uniform is in the same vein and presents no real difficulty. □

As in Theorem 1, we ignore in Theorem 2 constant multiplicative factors of the type $\lambda(1)$; we also ignore the coefficients $1/d_X$ of variables y or z. These factors might hide the global structure of Fig. 2, above, and do not serve any useful purpose: From §3.3, we know that the asymptotic study concerns the probability distribution generated by the function $f(x) = [y^r z^s]\Phi(x,y,z)/[y^r z^s]\Phi(1,y,z)$, and neither the elimination of a multiplicative factor in Φ nor the substitution of y for y/d_X have any effect on the function f. For example, the case detailed in the proof of Theorem 2, starting from the probability generating function for a uniform distribution, not from the counting generating function that we used, actually leads to the function $\Phi_0(x,y,z) = \left(e^{y/d_X} + e^{xy/d_X}(\lambda_S(z)-1)\right)^{d_X} \cdot \lambda_S(1)^{-d_X}$, and the function given in Fig. 2 is $\Phi(x,y,z) = (e^y + e^{xy}(\lambda_S(z)-1)^{d_X} = \lambda_S(1)^{d_X}\ \Phi_0(x, d_X y, z)$. Both functions lead to the same conditional probability distribution.

5.2. Limiting distributions. We can show that the semijoin size is asymptotically normal in several cases and that, as in the case of a projection, the probability distributions on attributes Y and U have almost no importance. There are three cases for each relation: It is free, or it has attribute X for key, or the other attribute is key (attribute Y for R, and attribute U for S). The choices for relations R and S are independent. As we see in §6.1, our method for proving results of asymptotic normality requires the evaluation of the coefficient $[y^r z^s]\Phi(x,y,z)$, and we classify the different cases according to the ease with which either one of the intermediate coefficients $[y^r]\Phi(x,y,z)$ or $[z^s]\Phi(x,y,z)$ can be computed. The possible cases for the two relations are listed below.

1. *None of relations R and S has X for key.* There is no direct way to evaluate $[y^r z^s]\Phi$, and we must use twice Cauchy's formula to compute it. We defer it to a future paper [10].
2. *R has X for key, but not S.* The extraction of $[y^r]\Phi$ gives a function in x and z: $\binom{d_X}{r}(1-x+x\lambda_S(z))^r \lambda_S(z)^{d_X - r}$. If, moreover, attribute U is key of relation S, then $\lambda_S(z)$ is the exponential function e^z. See Theorem 3 and Corollaries 2 and 3.
3. *S has X for key, but not R.* We compute the coefficient of z^s in the function Φ: $[z^s]\Phi = \binom{d_X}{s}\lambda_R(xy)^s \lambda_R(y)^{d_X - s}$. If relation R has attribute Y for key, then $\lambda_R(y) = e^y$. We then must study a bivariate function in x and y. This is done in Theorem 4 and Corollaries 4 and 5.
4. *R and S each have X for key.* The generating function has the simple form $\Phi(x,y,z) = (1+y+z+xyz)^{d_X}$, and either one of the coefficients $[y^r]\Phi$ and $[z^s]\Phi$ is easily computed. In this case, the semijoin of R and S has exactly the same number of elements as the intersection of relation R and the projection of relation S on attribute X, which has the same size as S. Conversely, the intersection of two relations can be seen as a semijoin of two relations with no other attribute than the one that is used in the join. Here again, we have a Gaussian limiting result (Theorem 5), which we state for the intersection of two relations built on the same attributes. Theorem 5 is also valid for the semijoin of two relations with a uniform probability distribution on the join attribute and without restriction on the distributions on the domains of

attributes Y and U.

The corollaries to the theorems below are valid for probability distributions in classes (Z) or (G) on the attributes Y and U, as indicated. In Theorems 3–5 and Corollaries 2–5, A and B are strictly positive constants. In the case of a probability distribution on either attribute Y or attribute U belonging to class (G), the domain sizes d_Y or d_U grow large (they are assumed to do so independently of each other and of d_X), and $\varphi_R(t)$ or $\varphi_S(t)$ denotes the limiting function $\prod_{i\geq 1}(1+p_i t)$; ρ is the unique real positive solution of the associated equation $t\varphi_R'(t)/\varphi_R(t) = A$ or $t\varphi_S'(t)/\varphi_S(t) = B$. The existence and uniqueness of ρ results from Lemma A in §3.4. The function g is defined by $g(t) = t\varphi_S'(t)/\varphi_S(t)$ in §5.3 and $g(t) = t\varphi_R'(t)/\varphi_R(t)$ in §5.4.

5.3. X key of R. We first give the general result pertaining to generating functions of the kind $\Phi(x,y,z) = (1+y+(1+xy)(\lambda_S(z)-1))^d$, then the corollaries dealing with the different cases for relation S. The proofs of Theorem 3 and of Corollaries 2 and 3 are given in §6.3.

THEOREM 3. *Let Property $\mathcal{P}$ be satisfied (see §3.4). Define $\Phi(x,y,z) = (1+y+(1+xy)(\lambda(z)-1))^d$. Let $d,r,s \to +\infty$ in such a way that $r < d$, $d = o(r^{3/2})$, and $s = Bd + o(d)$ for some positive constant B, and that $g(y) = y\lambda'(y)/\lambda(y)$ satisfies $\lim_{y\to+\infty} g(y) > B$. 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 Gaussian. The asymptotic values of the mean and variance are defined below in terms of the solution ρ of the equation $g(y) = B$:*

$$\mu = r\left(1 - \frac{1}{\lambda(\rho)}\right), \qquad \sigma^2 = r\left(\frac{\lambda(\rho)-1}{\lambda^2(\rho)} - \frac{r}{d}\frac{\rho\lambda'^2(\rho)}{\lambda^4(\rho)g'(\rho)}\right).$$

COROLLARY 2. *Let $R[X,Y]$ be a relation with a key X, and $S[X,U]$ a free relation. We assume that the probability distribution on D_X is uniform; the probability distribution on D_Y is arbitrary. The sizes r and s of the relations R and S are assumed to satisfy $r < d_X$, $d_X = o(r^{3/2})$, and $s = Bd_X(1+o(1))$. Then the probability distribution of the size of the semijoin of R and S on attribute X, conditioned by the sizes of R and S, is asymptotically normal.*

If the distribution on D_U satisfies hypothesis (Z), *the mean and variance have for asymptotic values $\mu = (1-e^{-B})r$ and $\sigma^2 = r((e^B-1)/e^{2B} - (rB)/d_X e^{2B})$. If the probability distribution on D_U satisfies hypothesis* (G), *the asymptotic values of the mean and variance are*

$$\mu = \mu_0 r = (1 - 1/\varphi_S(\rho))\, r, \qquad \sigma^2 = r\left(\frac{\varphi_S(\rho)-1}{\varphi_S^2(\rho)} - \frac{r}{d_X}\frac{\rho\varphi_S'^2(\rho)}{\varphi_S^4(\rho)g'(\rho)}\right).$$

Moreover, the constant μ_0 satisfies $1 - e^{-B} \leq \mu_0 \leq 1$.

COROLLARY 3. *Let $R[X,Y]$ be a relation with a key X, and $S[X,U]$ a relation with a key U. We assume that the probability distribution on D_X is uniform. The probability distributions on D_Y and D_U are arbitrary. The sizes r and s of the relations R and S are assumed to satisfy $r < d_X$, $d_X = o(r^{3/2})$, and $s = Bd_X(1+o(1))$. Then the probability distribution of the size of the semijoin of R and S on attribute X, conditioned by the sizes of R and S, is asymptotically normal. The mean and variance have for asymptotic values $\mu = (1-e^{-B})r$ and $\sigma^2 = r((e^B-1)/e^{2B} - rB/d_X e^{2B})$.*

The comparison of Corollary 2 in the case of a distribution of class (Z) on attribute U and of Corollary 3 shows that the two asymptotic distributions of the projection size are the same. As in the case of a projection, the exact distribution of the values of attribute U for a free relation, as long as it is not too far from uniform, or the existence of a key on U, are of no importance with respect to the asymptotic size of the semijoin when d_X and d_U go to infinity.

5.4. X key of S. This part deals with the cases where the generating function is of the kind $\Phi(x,y,z) = (\lambda_R(y) + z\lambda_R(xy))^d$. Here again, we first give the general result (Theorem 4), then the applications to the semijoin size (Corollaries 4 and 5), and we defer the proofs until §6.4.

THEOREM 4. *Let Property $\mathcal{P}$ be satisfied (see §3.4). Define $\Phi(x,y,z) = (\lambda(y) + z\lambda(xy))^d$. Let $d,r,s \to +\infty$ in such a way that $s < d$, $d = o(s^{3/2})$ and $r = Ad + o(d)$, and that $g(y) = y\lambda'(y)/\lambda(y)$ satisfies $\lim_{y\to+\infty} g(y) > A$. 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 Gaussian. The asymptotic value of the mean is $\mu = rs/d$. The asymptotic value of the variance is defined in terms of the solution ρ of the equation $g(y) = A$: $\sigma^2 = s(1-s/d)\rho g'(\rho)$.*

COROLLARY 4. *Let $R[X,Y]$ be a free relation, and $S[X,U]$ a relation with a key X. We assume that the probability distribution on D_X is uniform and that the probability distribution on D_Y is in class* (Z) *or* (G); *the probability distribution on D_U is arbitrary. The sizes r and s of the initial relations are assumed to satisfy $r = Ad_X + o(d_X)$, $s < d_X$, and $d_X = o(s^{3/2})$. Then the probability distribution of the size of the semijoin of R and S on attribute X, conditioned by the sizes r and s of the initial relations, is asymptotically normal. The asymptotic mean is $\mu = rs/d_X$. If the distribution on attribute Y belongs to class* (Z), *the asymptotic variance is equal to $\sigma^2 = (1-s/d_X)rs/d_X$. If the distribution on attribute Y belongs to class* (G), *the asymptotic variance becomes $\sigma^2 = s(1-s/d_X)\sigma_0^2$ for a positive constant σ_0^2.*

COROLLARY 5. *Let $R[X,Y]$ be a relation with key Y, and $S[X,U]$ a relation with a key X. We assume that the probability distribution on D_X is uniform. The probability distributions on D_Y and D_U are arbitrary. The sizes r and s of the initial relations are assumed to satisfy $r = Ad_X + o(d_X)$, $s < d_X$, and $d_X = o(s^{3/2})$. Then the probability distribution of the size of the semijoin of R and S on attribute X, conditioned by the sizes $r = Ad_X(1+o(1))$ of R and s of S, is asymptotically normal. The asymptotic mean and variance are given by $\mu = rs/d_X$ and $\sigma^2 = (1-s/d_X)rs/d_X$.*

Once again, a comparison of Corollaries 4 and 5 shows that the existence of a key, or a probability distribution of class (Z), on attribute Y have no influence on the asymptotic behaviour of the semijoin size.

In the case where both relations R and S have attribute X for key, we have the following result, which is proved in §6.5.

THEOREM 5. *Let R and S be two free relations, of sizes r and s. We assume that the probability distribution on the set of size d of possible tuples is uniform. We take $r = Ad(1+o(1))$ and $s = Bd(1+o(1))$, where A and B are constants in $]0,1[$. Then the probability distribution of the size of the intersection of R and S, conditioned by the sizes r and s of the initial relations, is asymptotically normal. The mean and variance are given by $\mu = rs/d$ and $\sigma^2 = rs/d(1-r/d)(1-s/d)$; their asymptotic values are, respectively, ABd and $AB(1-A)(1-B)d$.*

6. Proofs of theorems.

6.1. Sketch of the proofs. Theorems 1, 3–5 have a common flavor: We are interested in a function $\Phi(x,y)$ or $\Phi(x,y,z)$, which defines a conditional probability distribution, and we want to know if this distribution has a limit when some parameters r and s go to infinity. Moreover, the function Φ is of the kind ϕ^d, where the exponent d also grows to infinity. The generating function for the conditional distribution is $f(x) = [y^r]\Phi(x,y)/[y^r]\Phi(1,y)$ or $f(x) = [y^r z^s]\Phi(x,y,z)/[y^r z^s]\Phi(1,y,z)$.

Let us sketch the method that we use to study the limit of the distribution defined by $f(x)$ when function Φ is bivariate. When the initial function is $\Phi(x,y,z)$, we restrict ourselves in this paper to cases where at least one of the coefficients $[y^r]\Phi(x,y,z)$ or $[z^s]\Phi(x,y,z)$ can be computed by the binomial formula, and the evaluation of a limiting distribution defined by function $f(x)$ proceeds in a similar way.

We first evaluate $\psi(x) = [y^r]\Phi(x,y)$, for x real. By Cauchy's formula for an analytic function, $\psi(x)$ can be written as an integral on a closed contour around the origin as follows:

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

In all the cases that we consider in this paper, the function Φ has no singularity, and we use the saddlepoint method [3], [6], [14] to approximate this integral. We take for integration path in (2) a circle $y = \rho(x)e^{i\theta}$, centered at the origin, whose radius $\rho(x)$ is chosen in such a way that only a small part of the circle contributes to the integral and that the integral on the rest of the circle just gives an error term. The point $\rho(x)$ is a *saddlepoint*; it is defined from the function $h(x,y) = \log\Phi(x,y) - (r+1)\log y$ as the solution of the equation $(\partial h/\partial y)(x,y) = 0$. The saddlepoint approximation then gives the following approximate value of $\psi(x)$:

$$\psi(x) = \frac{e^{h(x,\rho(x))}}{\sqrt{2\pi\partial^2 h/\partial y^2(x,\rho(x))}}(1+o(1)).$$

We next show the pointwise convergence of the Laplace transform $e^{t\mu/\sigma}\psi(e^{-t/\sigma})/\psi(1)$ of the normalized random variable associated with the probability generating function $f(x) = \psi(x)/\psi(1)$, toward $e^{t^2/2}$, for suitably chosen values of μ and σ and for any fixed t in the interval $[0,+\infty[$. Classical results on the convergence of probability distributions (see, for example, [7, Chap. XIII, Thm. 2]) allow us to conclude to the convergence of the probability distribution defined by $f(x)$ toward a normal distribution of mean μ and variance σ^2.

We give in some detail the proof of Theorem 1 and, more quickly, the proofs of Theorems 3–5 and of the corollaries.

6.2. Proof of Theorem 1 for projections. We recall that the bivariate function we consider is $\Phi(x,y) = (1 - x + x\lambda(y))^d$, with Property $\mathcal{P}$ of §3.4 satisfied: It is entire and not affine, with positive coefficients, such that $\lambda(0) = 1$, and such that there exists no entire function Λ and integer $m \geq 2$ with $\lambda(y) = \Lambda(y^m)$. Let us define

$$h(x,y) = d\log(1 - x + x\lambda(y)) - (r+1)\log y. \tag{3}$$

Equation (2) becomes

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

6.2.1. Choice of the integration contour. The equation in y defining the saddlepoint is $\partial h/\partial y(x,y) = 0$, which can be rewritten as

$$\frac{xy\lambda'(y)}{1-x+x\lambda(y)} = \frac{r+1}{d}. \tag{4}$$

The solution of this equation in y is a function of x, r, and d. We must take x equal to 1, or near 1: $x = e^{-t/\sigma}$, for fixed t and large σ. We solve the equation for $x = 1$, then give an approximate solution for x close to and smaller than 1. For $x = 1$, (4) becomes

$$\frac{y\lambda'(y)}{\lambda(y)} = \frac{r+1}{d}. \tag{5}$$

We first show that (5) has a unique real positive solution ρ_0.

Lemma A of §3.4 and the fact that $g(0) = 0$ together show that ρ_0 exists and is unique if and only if $\lim_{y\to+\infty} y\lambda'(y)/\lambda(y) > (r+1)/d$, which can be simplified into a condition independent of d and r, below (we recall that $d, r \to +\infty$ and that $r = Ad + o(d)$):

$$\lim_{y\to+\infty} \frac{y\lambda'(y)}{\lambda(y)} > A. \tag{6}$$

We henceforth assume that the condition (6) holds: For r and d large enough, (5) has a unique real positive solution ρ_0. We look for a solution of (4) under the form $\rho(x) = (1+u)\rho_0$ with $u = o(1)$, for $x = 1+\varepsilon$ and $\varepsilon = o(1)$, with $\varepsilon < 0$.

Functions λ and λ' can be expanded near the origin as follows:

$$\lambda(y) = \lambda(\rho_0) + u\lambda'(\rho_0)\rho_0 + O(u^2),$$
$$\lambda'(y) = \lambda'(\rho_0) + u\lambda''(\rho_0)\rho_0 + O(u^2).$$

We rewrite (4) as $xy\lambda'(y)/(1-x+x\lambda(y)) = g(\rho_0)$ then plug the expansions of λ and λ' into it. In terms of the function

$$D(y) = \lambda(y)(\lambda'(y) + y\lambda''(y)) - y\lambda'^2(y),$$

this gives the following equation on u and $\varepsilon = x - 1$:

$$\lambda'(\rho_0)\varepsilon + D(\rho_0)u + O(u^2) + O(\varepsilon u) = 0.$$

The coefficient of u in this equation is $D(\rho_0) > 0$, and we can compute the following approximate value of u:

$$\rho(x) = (1+u)\rho_0; \qquad u = -\frac{\lambda'(\rho_0)}{D(\rho_0)}\varepsilon + O(\varepsilon^2). \tag{7}$$

6.2.2. Approximation of $\psi(x)$. In this section, x is fixed and real near 1. As we will need, in §6.2.3, to take $x = e^{-t/\sigma}$ for $t > 0$ and $\sigma > 0$, we can restrict ourselves to $x \le 1$. We also abbreviate $\partial^2 h/\partial y^2$ into h'' in this section; this should cause no ambiguity.

We choose for integration contour in the integral (2) a circle with radius $\rho(x)$:

$$\psi(x) = \frac{1}{2i\pi}\oint e^{h(x,y)}dy = \frac{1}{2i\pi}\int_{\theta\in[-\pi,+\pi]} e^{h(x,\rho(x)e^{i\theta})}d(\rho(x)e^{i\theta}).$$

We divide this integral in two parts: I_1 is the part of the integral dealing with a restricted piece of the path near point $\rho(x)$ and will give the main term (11); the complement I_2 will give an exponentially smaller term (12). We first choose $\alpha \in]0, \pi[$ (we see in the following the conditions that α must satisfy), then we formally define I_1 and I_2 by

$$I_1 = \frac{1}{2i\pi} \int_{\theta \in]-\alpha, +\alpha[} e^{h(x, \rho(x)e^{i\theta})} d(\rho(x)e^{i\theta}),$$

$$I_2 = \frac{1}{2i\pi} \int_{\alpha \le |\theta| \le \pi} e^{h(x, \rho(x)e^{i\theta})} d(\rho(x)e^{i\theta}).$$

We have that $\psi(x) = I_1 + I_2$, and our goal is to prove that we can find a value of α such that

$$\psi(x) = \frac{e^{h(x,\rho(x))}}{\sqrt{2\pi h''(x, \rho(x))}}(1 + o(1)). \tag{8}$$

Evaluation of I_1. We can always assume that x varies near 1 in such a way that $\rho(x)$ belongs to a compact set near ρ_0. Actually, ρ_0 itself is a function of d and r but can be restricted to a compact neighbourhood of $g^{-1}(A)$ (which is a constant). This means that we can restrict $\rho(x)$ to a compact interval around $g^{-1}(A)$ independently of r, d, and x. This will henceforth be used implicitly to prove that the error terms that we consider are uniform with respect to r, d, and x. We abbreviate $\rho(x)$ into ρ for the evaluation of I_1; again, this should cause no confusion. The evaluation of I_1 is similar to the corresponding ones in the proofs of Theorems 3 and 4. This leads us to state the following lemma, which we use again in the next proofs.

LEMMA B. *Let $h_d(x, y)$ be a function that depends on a parameter d, defined and twice differentiable for (x, y) in a compact neighbourhood of the point $(1, \rho_0)$, where ρ_0 satisfies the equation $\partial h_d/\partial y(1, \rho_0) = 0$. We assume furthermore that, as d varies, ρ_0 stays in a compact subset of $]0, +\infty[$. Define ρ as the solution (dependent on x and d) of $\partial h_d/\partial y(x, \rho) = 0$. Assume that*

- *$\partial^2 h_d/\partial y^2(1, \rho_0)$ is of order exactly d;*
- *for x near 1, $\partial^2 h_d/\partial y^2(x, \rho) = \partial^2 h_d/\partial y^2(1, \rho_0)(1 + o(1))$ with an error term uniform in x and independent of d;*
- *the function $\theta \mapsto h_d(x, \rho e^{i\theta})$ has a Taylor expansion near zero satisfying*

$$h_d(x, \rho e^{i\theta}) = h_d(x, \rho) + \frac{\rho^2}{2}(e^{i\theta} - 1)^2 \partial^2 h_d/\partial y^2(x, \rho) + O(d\theta^3),$$

with the $O()$ term such that the implied constant can be chosen independently of x and d.

Then there exist constants $\alpha_0 > 0$, γ_0, and γ_1 such that, for any $\alpha \in]0, \alpha_0[$, and with implied constants in the $O()$ terms independent of x and d,

$$\frac{1}{2i\pi} \int_{\theta \in]-\alpha, +\alpha[} e^{h_d(x, \rho e^{i\theta})} d(\rho e^{i\theta}) = \frac{e^{h_d(x,\rho)}}{\sqrt{2\pi \partial^2 h_d/\partial y^2(x, \rho)}} \cdot (1 + O(\alpha^2 \sqrt{d} e^{-\gamma_0 d\alpha^2}) + O(e^{-\gamma_1 d\alpha^2}) + O(d\alpha^3)).$$

The proof of Lemma B is given in Appendix A. We now check that its assumptions are satisfied for the function h defined by (3). The function $h(x, \rho e^{i\theta})$ can be expanded

in the variable θ around the origin as follows:

$$h(x,\rho e^{i\theta}) = h(x,\rho) + \rho(e^{i\theta}-1)\partial h/\partial y(x,\rho) + \frac{\rho^2}{2}(e^{i\theta}-1)^2 h''(x,\rho) + O(||h'''||(e^{i\theta}-1)^3). \tag{9}$$

Define $\phi(x,y) = 1-x+x\lambda(y)$. Definition (3) gives $h''(x,y) = d\cdot\partial^2(\log\phi)/\partial y^2(x,y) + (r+1)/y^2$. For x and y near 1 and ρ_0, respectively, we have that $\partial^2(\log\phi)/\partial y^2(x,y) = \partial^2(\log\phi)/\partial y^2(1,\rho_0) + O(x-1) + O(y-\rho_0)$ and $1/y^2 = 1/\rho_0^2 + O(y-\rho_0)$. This and the fact that $r = \Theta(d)$ give $h''(x,y) = h''(1,\rho_0) + d\,O(x-1) + d\,O(y-\rho_0)$. We next check that $h''(1,\rho_0)$ has order exactly d: $h''(1,\rho_0) = d\,g'(\rho_0)/\rho_0$. Hence we can factor it out of the expression of $h''(x,y)$, and we get that

$$h''(x,y) = h''(1,\rho_0)(1 + O(x-1) + O(y-\rho_0)). \tag{10}$$

The $O()$ terms in this expansion are independent of r and d. We deduce from it that, for x close to 1 and $y = \rho(x) = \rho$, $h''(x,\rho) = h''(1,\rho_0)(1+o(1))$. The error term in this relation is uniform for $x \to 1$ and $r, d \to +\infty$. A similar argument shows that the term $||h'''||$ is actually $O(d)$. We also have, by definition of ρ, that $\partial h/\partial y(x,\rho) = 0$. Equation (9) then becomes

$$h(x,\rho e^{i\theta}) = h(x,\rho) + \frac{\rho^2}{2}(e^{i\theta}-1)^2 h''(x,\rho) + O(d\theta^3).$$

Lemma B finally gives the following approximation of I_1:

$$I_1 = \frac{e^{h(x,\rho)}}{\sqrt{2\pi h''(x,\rho)}}(1 + O(\alpha^2\sqrt{d}e^{-\gamma_0 d\alpha^2})O(e^{-\gamma_1 d\alpha^2}) + O(d\alpha^3)). \tag{11}$$

Upper bound on I_2. We recall that

$$I_2 = \frac{1}{2i\pi}\int_{\alpha\le|\theta|\le\pi} e^{h(x,\rho(x)e^{i\theta})} d(\rho(x)e^{i\theta}).$$

We extract from the integral the main term of I_1: $e^{h(x,\rho(x))}$; this gives

$$I_2 = \frac{\rho(x)e^{h(x,\rho(x))}}{2\pi}\int_{\alpha\le|\theta|\le\pi} e^{-ir\theta}k_x(\theta)^d d\theta,$$

with $k_x(\theta) = (1-x+x\lambda(\rho(x)e^{i\theta}))/(1-x+x\lambda(\rho(x)))$. We now want an upper bound on $|k_x(\theta)|$, for $|\theta| \in [\alpha,\pi]$. The term $1-x$ is $o(1)$, of smaller order than the term $\lambda(\rho(x))$, and we get that

$$|k_x(\theta)| \le \frac{1-x+x|\lambda(\rho(x)e^{i\theta})|}{1-x+x\lambda(\rho(x))}.$$

Lemma C, below (proved in Appendix B), gives a bound on $|\lambda(\rho(x)e^{i\theta})|$, which, in turn, gives an inequality on $|k_x(\theta)|$: $|k_x(\theta)| \le 1-C_1\alpha^2$, for a strictly positive constant C_1, independent of r, d, x, and α. We finally get that

$$|I_2| \le \rho(x)e^{h(x,\rho(x))}(1-C_1\alpha^2)^d = e^{h(x,\rho(x))}O(e^{-C_1 d\alpha^2}),$$

which gives

$$|I_2| = \frac{e^{h(x,\rho(x))}}{\sqrt{h''(x,\rho(x))}} O(\sqrt{d}\, e^{-C_1 d\alpha^2}). \tag{12}$$

LEMMA C. *Let λ be a function satisfying Property $\mathcal{P}$ of §3.4, and $\alpha \in]0,\pi[$. Let y vary in a compact subset of $]0,+\infty[$. Then there exists a constant $C > 0$ such that, for all θ satisfying $\alpha < |\theta| < \pi$, and for all y in the compact subset, the following inequality holds:*

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

Choice of α. We obtain the following approximation of $\psi(x)$:

$$\psi(x) = \frac{e^{h(x,\rho(x))}}{\sqrt{2\pi h''(x,\rho(x))}}(1+O(e^{-\gamma_1 d\alpha^2})+O(\sqrt{d}e^{-C_1 d\alpha^2})+O(\alpha^2\sqrt{d}e^{-\gamma_0 d\alpha^2})+O(d\alpha^3)).$$

Approximation (8) holds if we can choose α such that the error terms are negligible. For $d \to +\infty$, this is a consequence of

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

For $\alpha = (\log d)/\sqrt{d}$, it is easy to check that $d\alpha^2/\log d = \log d$ and $d\alpha^3 = \log^3 d/\sqrt{d} = o(1)$, and we have that

$$\psi(x) = \frac{e^{h(x,\rho(x))}}{\sqrt{2\pi h''(x,\rho(x))}}(1 + o(1)).$$

6.2.3. Convergence of the normalized Laplace transform. We show here that we can choose μ and σ in such a way that the function $e^{t\mu/\sigma} f(e^{-t/\sigma}) = e^{t\mu/\sigma}\psi(e^{-t/\sigma})/\psi(1)$ converges toward $e^{t^2/2}$ when $d \to +\infty$ and for every $t > 0$. Taking the logarithm, we must prove the convergence of $\Xi(t) = t\mu/\sigma + \log(\psi(e^{-t/\sigma})/\psi(1))$ toward $t^2/2$.

Equation (8) shows that

$$\log\left(\frac{\psi(x)}{\psi(1)}\right) = h(x,\rho(x)) - h(1,\rho_0) - \frac{1}{2}\log\left(\frac{h''(x,\rho(x))}{h''(1,\rho_0)}\right) + o(1),$$

which gives

$$\Xi(t) = t\mu/\sigma + \delta h(e^{-t/\sigma}, \rho(e^{-t/\sigma})) + o(1), \tag{13}$$

with

$$\delta h(x,y) = h(x,y) - h(1,\rho_0) - \frac{1}{2}\log\frac{h''(x,y)}{h''(1,\rho_0)}. \tag{14}$$

We formerly proved in (10) that h'' can be expanded near the point $(1,\rho_0)$, $h''(x,y) = h''(1,\rho_0)(1 + O(x-1) + O(y-\rho_0))$. This shows that, for $x = 1+\varepsilon$,

$$\log\left(\frac{h''(x,\rho(x))}{h''(1,\rho_0)}\right) = O(\varepsilon).$$

Evaluation of $h(x,\rho(x)) - h(1,\rho_0)$. The function h can be expanded near the point $(1,\rho_0)$ as follows:

$$h(x,y) = h(1,\rho_0) + (x-1)\frac{\partial h}{\partial x}(1,\rho_0) + (y-\rho_0)\frac{\partial h}{\partial y}(1,\rho_0)$$
$$+\frac{1}{2}(x-1)^2\frac{\partial^2 h}{\partial x^2}(1,\rho_0) + (x-1)(y-\rho_0)\frac{\partial^2 h}{\partial x \partial y}(1,\rho_0) + \frac{1}{2}(y-\rho_0)^2\frac{\partial^2 h}{\partial y^2}(1,\rho_0)$$
$$+O(d(x-1)^3) + O(d(y-\rho_0)^3).$$

For $x-1=\varepsilon$ and $y-\rho_0 = u\rho_0$, and substituting the values of the derivatives of h at point $(1,\rho_0)$, we get that

$$h(1+\varepsilon,(1+u)\rho_0) = h(1,\rho_0) + d\frac{\lambda(\rho_0)-1}{\lambda(\rho_0)}\varepsilon - \frac{d}{2}\cdot\frac{(\lambda(\rho_0)-1)^2}{\lambda^2(\rho_0)}\varepsilon^2$$
$$+d\frac{\lambda'(\rho_0)}{\lambda(\rho_0)}\varepsilon u\rho_0 + \frac{d}{2}\,g'(\rho_0)u^2\rho_0 + O(d\varepsilon^3).$$

We now use the value of u computed in (7): $u = -\varepsilon\lambda'(\rho_0)/D(\rho_0) + O(\varepsilon^2)$, which gives

$$h(1+\varepsilon,\rho_0(1+u)) = h(1,\rho_0) + d\alpha_1\varepsilon - \frac{d}{2}\alpha_2\varepsilon^2 + O(d\varepsilon^3).$$

The coefficients α_1 and α_2 in this formula are defined by

$$\alpha_1 = 1 - \frac{1}{\lambda(\rho_0)}, \qquad \alpha_2 = \alpha_1^2 + \frac{\lambda'^2(\rho_0)\rho_0}{\lambda^2(\rho_0)D(\rho_0)}.$$

The values of h and h'' in (14) are now replaced and we get that

$$\delta h(1+\varepsilon,(1+\varepsilon)\rho_0) = d\alpha_1\varepsilon - \frac{d}{2}\alpha_2\varepsilon^2 + O(d\varepsilon^3) + O(\varepsilon). \tag{15}$$

We next define $x = e^{-t/\sigma} = 1 - t/\sigma + t^2/2\sigma^2 + O(t^3/\sigma^3)$, i.e., $\varepsilon = -t/\sigma + t^2/2\sigma^2 + O(t^3/\sigma^3)$. Equations (13) and (15) show that

$$\Xi(t) = (\mu - d\alpha_1)\frac{t}{\sigma} + d(\alpha_1-\alpha_2)\frac{t^2}{2\sigma^2} + O\left(\frac{d}{\sigma^3}\right) + O\left(\frac{1}{\sigma}\right) + o(1).$$

Determination of the mean μ *and the variance* σ. Let us define $\mu = d\alpha_1$ and $\sigma^2 = d(\alpha_1-\alpha_2)$. They can be written as

$$\mu = d\frac{\lambda(\rho_0)-1}{\lambda(\rho_0)}, \qquad \sigma^2 = d\left(\frac{\lambda(\rho_0)-1}{\lambda^2(\rho_0)} - \frac{\lambda'^2(\rho_0)\rho_0}{\lambda^2(\rho_0)D(\rho_0)}\right).$$

Here ρ_0 is equal to $g^{-1}(r+1)/d$ and has for asymptotic value the solution of $g(y) = A$. Hence the mean and variance are asymptotically equal, respectively, to $d\mu_0$ and $d\sigma_0^2$, for μ_0 and σ_0 defined in function of $\rho = g^{-1}(A)$ as follows:

$$\mu_0 = \frac{\lambda(\rho)-1}{\lambda(\rho)}, \qquad \sigma_0^2 = \frac{\lambda(\rho)-1}{\lambda^2(\rho)} - \frac{\rho\lambda'^2(\rho)}{\lambda^2(\rho)D(\rho)}.$$

We now check that σ_0^2 is strictly positive. In terms of the functions $g_1(y) = y\lambda'(y)/(\lambda(y) - 1)$ and $g(y) = y\lambda'(y)/\lambda(y)$, we have that $\sigma_0^2 = (\lambda(\rho) - 1)^2 g_1'(\rho)/(\lambda^3(\rho) g'(\rho))$. By an argument similar to that used in §3.4 to prove that g is an increasing function (see Lemma A and its proof), we can show that the value $g_1'(\rho)$ is positive, which in turn shows that $\sigma_0^2 > 0$. The error terms $O(d/\sigma^3)$ and $O(1/\sigma)$ both become $O(1/\sqrt{d}) = o(1)$, and we finally have that $\Xi(t) = t^2/2 + o(1)$, which ends the proof of Theorem 1. □

6.2.4. Proof of Corollary 1. Checking that the function $\lambda_{d_Y}(y) = \prod_{1 \le i \le d_Y}(1 + p_{i,d_Y} y)$ satisfies Property $\mathcal{P}$ presents no difficulty. If the size d_Y of D_Y is fixed, Corollary 1 is a direct consequence of Theorem 1. If d_Y grows to infinity independently of d_X and r, we must adjust the proof as indicated below. We recall that we assume the independence of d_X and d_Y.

We work with a *sequence* of functions $\lambda_{d_Y}(y) = \prod_{1 \le i \le d_Y}(1 + p_{i,d_Y} y)$. When the probability distribution on attribute Y is in class (Z) or (G), this sequence converges normally toward a function $\varphi(y)$ for any y in a compact subset of the complex plane and for $d_Y \to +\infty$. The saddlepoint ρ_0 for $x = 1$ has a finite, nonnull limit ρ when $r, d_X, d_Y \to +\infty$. This limit ρ also satisfies the limiting equation $t\varphi'(t)/\varphi(t) = A$. We solve the equation, giving the general saddlepoint $\rho(x)$ exactly as in §6.2.1. The solution now also depends on d_Y, and it is important to note that $\rho(x)$ can be restricted to a compact neighbourhood of ρ for $x \to 1$ and $r, d_X, d_Y \to +\infty$. The rest of the proof is then the same as the corresponding part of the proof of Theorem 1, with uniform error terms in our approximations.

When the distribution on attribute Y belongs to class (G), the inequality $\mu_0 \ge 1 - e^{-A}$ is equivalent to $e^A \le \varphi(\rho)$ or (by $g(\rho) = A$) to $g(\rho) < \log \phi(\rho)$. As $g(y) = y(\log \phi)'(y)$ and $\log \varphi(y) = \sum_{i \ge 1} \log(1 + p_i y)$, we have that $g(y) = \sum_{i \ge 1} p_i y/(1 + p_i y)$. We can then rewrite the former inequality as $\sum_{i \ge 1}(\log(1 + p_i\rho) - p_i\rho/(1 + p_i\rho)) \ge 0$. The function $t \mapsto \log(1 + t) - t/(1 + t)$ is positive for all $t \in]0, +\infty[$, and each of the terms of the global inequality is positive, which proves the lower bound on μ_0. □

6.3. Proof of Theorem 3 for semijoins: X key of R. Theorem 3 is an extension of Theorem 1, when we multiply the function $\Phi(x, y) = (1 - x + x\lambda(y))^d$ by a term $\varphi(z) = \lambda(z)^{d-r}$. The proof of Theorem 3 is similar to that of Theorem 1, and we mainly indicate the points where it differs.

The coefficient $[y^r]\Phi(x, y, z)$ is $\binom{d}{r}(1 - x + x\lambda(z))^r \lambda(z)^{d-r}$. Let us define

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

with

$$h(x, z) = r \log(1 - x + x\lambda(z)) + (d - r) \log \lambda(z) - (s + 1) \log z.$$

6.3.1. Evaluation of the saddlepoint $z(x)$. We have that $h(1, z) = d \log \lambda(z) - (s + 1) \log z$. Define $g(z) = z\lambda'(z)/\lambda(z)$; the equation $\partial h/\partial z(1, z) = 0$ becomes $g(z) = (s + 1)/d$. We assume that $\lim_{s,d \to +\infty}(s + 1)/d$ exists and is equal to B. As function λ satisfies Property $\mathcal{P}$ of §3.4, Lemma A of that section shows that the equation $g(z) = (s + 1)/d$ has a unique real positive solution ρ_0 if and only if $\lim_{z \to +\infty} g(z) > B$. We then solve the equation $\partial h/\partial z(1, z) = 0$, for $x = 1 + \varepsilon$ and $z = (1 + u)\rho_0$. We first rewrite it into

$$\left(1 + \frac{r(x - 1)}{d(1 - x + x\lambda(z))}\right) g(z) = \frac{s + 1}{d}.$$

Using expansions of the functions λ and g near ρ_0, we get the approximate equation

$$\frac{rg(\rho_0)}{d\lambda(\rho_0)}\varepsilon + g'(\rho_0)u\rho_0 + O\left(\frac{r}{d}\varepsilon^2\right) + O\left(\frac{r}{d}\varepsilon u\right) + O(u^2) = 0.$$

We solve it and get that

$$u = -\alpha\varepsilon(1+O(\varepsilon)), \qquad \alpha = \frac{r\lambda'(\rho_0)}{d\lambda^2(\rho_0)g'(\rho_0)}. \tag{16}$$

We need the values of the derivatives of function h near point $(1,\rho_0)$ in §§6.3.2 and 6.3.3, so we give them below: $\partial h/\partial z(1,\rho_0) = 0$, the derivatives of order 3 of h are $O(d)$, and

$$\frac{\partial h}{\partial x}(1,\rho_0) = r\left(1-\frac{1}{\lambda(\rho_0)}\right), \qquad \frac{\partial^2 h}{\partial x^2}(1,\rho_0) = -r\left(1-\frac{1}{\lambda(\rho_0)}\right)^2,$$
$$\frac{\partial^2 h}{\partial x\partial z}(1,\rho_0) = r\,\frac{\lambda'(\rho_0)}{\lambda^2(\rho_0)}, \qquad \frac{\partial^2 h}{\partial z^2}(1,\rho_0) = d\,\frac{g'(\rho_0)}{\rho_0}.$$

6.3.2. Evaluation of $\psi(x)$. In the formula $\psi(x) = (1/2i\pi)\oint e^{h(x,z)}dz$, we take for integration path a circle centered at the origin and with radius $z(x) = (1+u)\rho_0$, with u defined by (16). We choose $\alpha \in]0,\pi[$ and divide the integral in two parts: $I_1 = (1/2i\pi)\int_{|\theta|\le\alpha} e^{h(x,z)}dz$ and $I_2 = (1/2i\pi)\int_{\alpha\le|\theta|\le\pi} e^{h(x,z)}dz$. Lemma B of §6.2.2 gives an approximation of I_1 as follows:

$$I_1 = \frac{e^{h(x,z(x))}}{\sqrt{2\pi h''_{z^2}(x,z(x))}}(1+O(\alpha^2\sqrt{d}e^{-\gamma_0 d\alpha^2}) + O(e^{-\gamma_1 d\alpha^2}) + O(d\alpha^3)).$$

Lemma C of §6.2.2 then gives an upper bound on $|\lambda(z(x)e^{i\theta})|$ for $\alpha \le |\theta| \le \pi$; it is easy to show from it that

$$I_2 = \frac{e^{h(x,z(x))}}{\sqrt{h''_{z^2}(x,z(x))}}\,O(\sqrt{d}e^{-\gamma_2 d\alpha^2}).$$

By choosing $\alpha = (\log d)/\sqrt{d}$, we obtain that

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

6.3.3. Laplace transform and determination of moments. Always following the same path as in the proof of Theorem 1, we now compute

$$\log\frac{\psi(x)}{\psi(1)} = h(x,z(x)) - h(1,\rho_0) - \frac{1}{2}\log\frac{\partial^2 h/\partial z^2(x,z(x))}{\partial^2 h/\partial z^2(1,\rho_0)} + o(1).$$

It is easy to check that $\log(\partial^2 h/\partial z^2(x,z(x))/\partial^2 h/\partial z^2(1,\rho_0))$ is $O(x-1)$. We then expand the function $h(x,z(x))$ near the point $(1,\rho_0)$ as follows:

$$\begin{aligned} h(x,z(x)) = h(1,\rho_0) &+ (x-1)\frac{\partial h}{\partial x} + (z(x)-\rho_0)\frac{\partial h}{\partial z} + \frac{1}{2}(x-1)^2\frac{\partial^2 h}{\partial x^2} \\ &+ (x-1)(z(x)-\rho_0)\frac{\partial^2 h}{\partial x\partial z} + \frac{1}{2}(z(x)-\rho_0)^2\frac{\partial^2 h}{\partial z^2} \\ &+ O(d(x-1)^3) + O(d(z(x)-\rho_0)^3). \end{aligned}$$

The values of the derivatives of h in this expansion are taken at point $(1, \rho_0)$. For $x = 1 + \varepsilon$ and $z(x) = (1 + u)\rho_0$ (see (16)), we get that

$$h(x, z(x)) - h(1, \rho_0) = \frac{\partial h}{\partial x}\varepsilon + \frac{1}{2}\left(\frac{\partial^2 h}{\partial x^2} - 2\alpha\rho_0\frac{\partial^2 h}{\partial x \partial z} + \alpha^2\rho_0^2\frac{\partial^2 h}{\partial z^2}\right)\varepsilon^2 + O(d\varepsilon^3).$$

Define $\mu = \partial h/\partial x$ and $\sigma^2 = \mu + \partial^2 h/\partial x^2 - 2\alpha\rho_0\partial^2 h/\partial x\partial z + \alpha^2\rho_0\partial^2 h/\partial z^2$; the values of the derivatives in μ and σ^2 are taken at point $(1, \rho_0)$. We have that

$$\log\frac{\psi(x)}{\psi(1)} = \mu(x-1) + \frac{1}{2}(\sigma^2 - \mu)(x-1)^2 + O(d(x-1)^3) + O(x-1).$$

For $s, d_X \to +\infty$, we have that $\mu = r\mu_0(1 + o(1))$ and $\sigma^2 = r\sigma_0^2(1 + o(1))$, with the constants μ_0 and σ_0 defined in function of $\rho = g^{-1}(B)$ as follows:

$$\mu_0 = 1 - \frac{1}{\lambda(\rho)}, \qquad \sigma_0^2 = \frac{\lambda(\rho) - 1}{\lambda^2(\rho)} - \frac{r}{d}\frac{\rho\lambda'^2(\rho)}{\lambda^4(\rho)g'(\rho)}.$$

We again note that σ_0^2 is strictly positive: $\sigma_0^2 \geq (\lambda(\rho)-1)/\lambda^2(\rho) - \rho\lambda'^2(\rho)/\lambda^4(\rho)g'(\rho)$, and we proved in §6.2.3 that this term is strictly positive. We finally get that $t\mu/\sigma + \log\psi(e^{-t/\sigma})/\psi(1) = t^2/2 + O(d/r^{3/2}) + O(1/r) + o(1)$. The error term becomes $o(1)$ for r such that $r \to +\infty$ and $r^{3/2}/d \to +\infty$. □

6.3.4. Proofs of Corollaries 2 and 3. The proof of Corollary 2 is adapted from the proof of Theorem 3, as Corollary 1 was obtained from Theorem 1 in §6.2.4: Take a sequence of functions $\lambda_{d_Z}(t)$ and note that $z(x)$ can be restricted to a compact subset near $g^{-1}(B)$. Corollary 3 is simply Theorem 3 applied to the function $\lambda(z) = e^z$.

6.4. Proof of Theorem 4 for semijoins: X key of S. The generating function $\Phi(x, y, z)$ has the following general form:

$$\Phi(x, y, z) = (\lambda(y) + z\lambda(xy))^d.$$

We first extract the coefficient of z^s in $\Phi(x, y, z)$ as follows:

$$[z^s]\Phi(x, y, z) = \binom{d}{s}\lambda(xy)^s\lambda(y)^{d-s}.$$

Cauchy's formula then gives the following coefficient of y^r:

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

with

$$h(x, y) = (d - s)\log\lambda(y) + s\log\lambda(xy) - (r + 1)\log y.$$

Here again, we choose for integration path a circle centered at the origin and that has for radius the root $y(x)$ of equation $\partial h/\partial y(x, y) = 0$.

6.4.1. Evaluation of the saddlepoint $y(x)$. For $x = 1$, $y(x)$ is solution of $\partial h/\partial y(1, y) = 0$. We have that $h(1, y) = d \log \lambda(y) - (r + 1) \log y$. This gives

$$y\frac{\lambda'(y)}{\lambda(y)} = \frac{r+1}{d}. \tag{17}$$

Define again $g(y) = y\lambda'(y)/\lambda(y)$. Equation (17) can be written as $g(y) = (r + 1)/d$. As function λ satisfies Property $\mathcal{P}$, the equation $g(y) = (r+1)/d$ has a unique solution ρ_0 if and only if (cf. Lemma A in §3.4)

$$\lim_{y\to+\infty} y\frac{\lambda'}{\lambda}(y) > A.$$

We then solve (17) for $x = 1 + \varepsilon$ and $y(x) = (1 + u)\rho_0$. They satisfy the equation

$$(d - s)g(y) + s\, g(xy) - (r + 1) = 0. \tag{18}$$

Function g can be expanded near ρ_0, as follows:

$$g(y) = g(\rho_0) + (y - \rho_0)g'(\rho_0) + O(||g''||(y - \rho_0)^2).$$

The error term is simply $O((y - \rho_0)^2)$, and we have for $y = (1 + u)\rho_0$ that

$$g(y) = g(\rho_0) + g'(\rho_0)u\rho_0 + O(u^2).$$

As $xy = (1 + \varepsilon + u + \varepsilon u)\rho_0$, we get that

$$g(xy) = g(\rho_0) + g'(\rho_0)(\varepsilon + u)\rho_0 + O(\varepsilon^2) + O(u^2).$$

Equation (18) can be simplified by using $r + 1 = dg(\rho_0)$, and we get the following approximate equation between ε and u:

$$du + s\varepsilon + O(du^2) + O(s\varepsilon^2) = 0.$$

We have that $s < d$, and we can solve this equation in u. This gives the following approximate value of the saddlepoint for $x = 1 + \varepsilon$:

$$y(x) = (1 + u)\rho_0, \qquad u = -\frac{s}{d}\varepsilon(1 + O(\varepsilon)). \tag{19}$$

We indicate below the values of derivatives of h that we need later: $\partial h/\partial y(1, \rho_0) = 0$, the derivatives of order 3 of h near $(1, \rho_0)$ are $O(d)$, and

$$\frac{\partial h}{\partial x}(1, \rho_0) = sg(\rho_0), \qquad \frac{\partial^2 h}{\partial x^2}(1, \rho_0) = s(\frac{\lambda'}{\lambda})'(\rho_0)\rho_0^2 = s(g'(\rho_0)\rho_0 - g(\rho_0)),$$
$$\frac{\partial^2 h}{\partial x \partial y}(1, \rho_0) = sg'(\rho_0), \qquad \frac{\partial^2 h}{\partial y^2}(1, \rho_0) = d\frac{g'(\rho_0)}{\rho_0}.$$

6.4.2. Approximation of $\psi(x)$. We take here x fixed, real, and smaller than 1. The function $\psi(x) = [y^r z^s]\Phi(x, y, z)/\binom{d}{s}$ can be written as an integral along a circle of center the origin and radius $y(x) = (1 + u)\rho_0$, below:

$$\psi(x) = \frac{1}{2i\pi}\oint e^{h(x,y)}dy = \frac{1}{2i\pi}\int_{\theta\in[-\pi,+\pi]} e^{h(x,y(x)e^{i\theta})}d(y(x)e^{i\theta}).$$

For α in $]0,\pi[$, we define

$$I_1 = \frac{1}{2i\pi}\int_{\theta\in]-\alpha,+\alpha[} e^{h(x,y(x)e^{i\theta})}d(y(x)e^{i\theta}),$$
$$I_2 = \frac{1}{2i\pi}\int_{\alpha\le|\theta|\le\pi} e^{h(x,y(x)e^{i\theta})}d(y(x)e^{i\theta}).$$

Evaluation of I_1. We check that the assumptions of Lemma B of §6.2.2 are satisfied: The conditions on the derivatives of h hold, and $h(x,y)$ has the following expansion for y near the saddlepoint $y(x)$:

$$h(x,y) = h(x,y(x)) + (y-y(x))\frac{\partial h}{\partial y}(x,y(x)) + 1/2\,(y-y(x))^2\frac{\partial^2 h}{\partial y^2}(x,y(x))$$
$$+O(||h'''||(y-y(x))^3).$$

This gives

$$h(x,y(x)e^{i\theta}) = h(x,y(x)) + \frac{y(x)^2}{2}(e^{i\theta}-1)^2 h''_{y^2}(x,y(x)) + O(d\theta^3).$$

Lemma B then proves that

(20) $$I_1 = \frac{e^{h(x,y(x))}}{\sqrt{2\pi h''_{y^2}(x,y(x))}}(1+O(\alpha^2\sqrt{d}e^{-\gamma_0 d\alpha^2}) + O(e^{-\gamma_1 d\alpha^2}) + O(d\alpha^3)).$$

Upper bound on I_2. We have that

$$I_2 = \frac{e^{h(x,y(x))}}{2i\pi}\int_{\alpha\le|\theta|\le\pi}\left(\frac{\lambda(y(x)e^{i\theta})}{\lambda(y(x))}\right)^{d-s}\cdot\left(\frac{\lambda(xy(x)e^{i\theta})}{\lambda(y(x))}\right)^{s}\cdot e^{-i(r+1)\theta}d\theta.$$

Lemma C in §6.2.2 shows that there exists a suitable constant $\gamma > 0$ such that the following inequalities hold:

$$|\lambda(y(x)e^{i\theta})| \le \lambda(y(x))e^{-\gamma\alpha^2}, \qquad |\lambda(xy(x)e^{i\theta})| \le \lambda(xy(x))e^{-\gamma\alpha^2}.$$

As a consequence,

$$\int_{\alpha\le|\theta|\le\pi}\left|\frac{\lambda(y(x)e^{i\theta})}{\lambda(y(x))}\right|^{d-s}\cdot\left|\frac{\lambda(xy(x)e^{i\theta})}{\lambda(y(x))}\right|^{s}\cdot d\theta \le 2\pi e^{-\gamma d\alpha^2}.$$

We get that

(21) $$I_2 = \frac{e^{h(x,y(x))}}{\sqrt{2\pi h''_{y^2}(x,y(x))}}O(\sqrt{d}e^{-\gamma d\alpha^2}).$$

Choice of α. As usual, we choose $\alpha = (\log d)/\sqrt{d}$; the error terms in (20) and (21) then become $o(1)$ for $r,d\to+\infty$. Hence

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

6.4.3. Convergence of the Laplace transform. We show here that $e^{t\mu/\sigma}\psi(e^{-t/\sigma})/\psi(1)$ converges toward $e^{t^2/2}$ for $d \to +\infty$ and for all fixed real positive t. Its logarithm is $\Xi(t) = t\mu/\sigma + \log(\psi(e^{-t/\sigma})/\psi(1))$. For $x = 1+\varepsilon$ and $y = (1+u)\rho_0$, and using the information on the order of the derivatives of h, we get that

$$h(x,y) = h(1,\rho_0) + \frac{\partial h}{\partial x}(1,\rho_0)\varepsilon + \frac{\partial h}{\partial y}(1,\rho_0)u\rho_0 + \frac{1}{2}\frac{\partial^2 h}{\partial x^2}(1,\rho_0)\varepsilon^2 + \frac{\partial^2 h}{\partial x \partial y}(1,\rho_0)\varepsilon u\rho_0 + \frac{1}{2}\frac{\partial^2 h}{\partial y^2}(1,\rho_0)u^2\rho_0^2 + O(du^3) + O(d\varepsilon^3).$$

We substitute $-s\varepsilon/d \cdot (1+O(\varepsilon))$ for u (see (19)), and the values computed above for the derivatives of h, and we get that

$$h(1+\varepsilon, \rho_0(1+u)) - h(1,\rho_0) = sg(\rho_0)\varepsilon + \frac{s}{2}\left(\left(1-\frac{s}{d}\right)g'(\rho_0)\rho_0 - g(\rho_0)\right)\varepsilon^2 + O(d\varepsilon^3).$$

We have, as usual, that $(\partial^2 h/\partial y^2(x,\rho_0(x)))/(\partial^2 h/\partial y^2(1,\rho_0)) = O(\varepsilon)$. For $x = e^{-t/\sigma}$, we then substitute $-t/\sigma + t^2/2\sigma^2 + O(1/\sigma^3)$ for $\varepsilon = x-1$, and we obtain that

$$\Xi(t) = (\mu - sg(\rho_0))\frac{t}{\sigma} + s(1-s/d)g'(\rho_0)\rho_0\frac{t}{2\sigma^2} + O\left(\frac{d}{\sigma^3}\right) + O\left(\frac{1}{\sigma}\right).$$

Define $\mu = sg(\rho) = As \approx rs/d$ and $\sigma^2 = s(1-s/d)\rho g'(\rho)$, with $\rho = g^{-1}(A)$. The conditions on s and d show that the error terms are $o(1)$ and we have that $\Xi(t) \to e^{t^2/2}$. □

6.4.4. Proofs of Corollaries 4 and 5. The proof of Corollary 4 is adapted from that of Theorem 4, as indicated in §6.2.4 for Corollary 1 and Theorem 1. Corollary 5 is an instance of Theorem 4 in the case when $\lambda(y) = e^y$.

6.5. Proof of Theorem 5. Theorem 5 cannot be deduced from either Theorem 3 or Theorem 4: The functions $\lambda_R(t)$ and $\lambda_S(t)$ are both equal to $1+t$. However, the function $\Phi(x,y,z)$ is simple enough that it is possible to write a direct proof. We can express $[y^r z^s]\Phi(x,y,z)$ as a sum of binomial coefficients: For $\Phi(x,y,z) = (1+y+z+xyz)^d$, we have that

$$[y^r z^s]\Phi(x,y,z) = \sum_{i+j=s}\binom{d}{r}\binom{d-r}{i}\binom{r}{j}x^j.$$

We can then try a direct study based on properties of the binomial coefficients. We do not follow this idea, but rather indicate briefly how Theorem 5 can be proved by our approach.

We compute $[y^r]\Phi$, then apply Cauchy's formula to get $[y^r z^s]\Phi$ as follows:

$$[y^r z^s]\Phi(x,y,z) = \frac{\binom{d}{r}}{2i\pi}\oint e^{h(x,z)}dz,$$

with $h(x,z) = (d-r)\log(1+z) + r\log(1+xz) - (s+1)\log z$.

The saddlepoint for $x=1$ is $\rho_0 = (s+1)/(d-s-1)$. We must assume that $s = Bd + o(d)$ if we want ρ_0 to stay in a compact subset of $]0,+\infty[$. For $x = 1+\varepsilon$, the saddlepoint is $z(x) = \rho_0(1 - r\varepsilon/d + O(r\varepsilon/d))$.

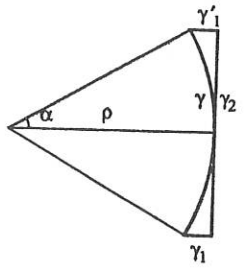
As usual, the computation of the integral $\int_{[-\pi,+\pi]} e^{h(x,z(x)e^{i\theta}}d(z(x)e^{i\theta})$ is performed in two parts. The computation of the main part, on interval $[-\alpha,+\alpha]$, is

straightforward, and we do not detail it. The upper bound on the remainder of the integral once again relies on the bound on a function of θ, for $\alpha \leq |\theta| \leq \pi$. In the present case, this function is simply $(1+z(x)e^{i\theta})^{d-r}(1+xz(x)e^{i\theta})^r/(1+z(x))^d$, and the desired inequality presents no difficulty. The evaluation of the normalized Laplace transform and its convergence toward $e^{t^2/2}$ are then easily proved. □

Appendix A. Proof of Lemma B. We write h instead of h_d, and h'' instead of $\partial^2 h_d/\partial y^2$. Let $\alpha_0 > 0$ be such that the Taylor expansion of h is valid for $|\theta| \leq \alpha_0$. Then, for any $\alpha \leq \alpha_0$, the assumptions on h show that

$$\frac{1}{2i\pi}\int_{\theta\in]-\alpha,+\alpha[} e^{h(x,\rho e^{i\theta})}d(\rho e^{i\theta}) = \frac{e^{h(x,\rho)}}{2i\pi}\int_{|\theta|<\alpha} e^{(\rho^2/2)(e^{i\theta}-1)^2 h''(x,\rho)+O(d\theta^3)}d(\rho e^{i\theta})$$
$$= \frac{e^{h(x,\rho)}}{2i\pi}\int_{|\theta|<\alpha} e^{(\rho^2/2)(e^{i\theta}-1)^2 h''(x,\rho)}d(\rho e^{i\theta})\ (1+O(d\alpha^3)).$$

Let us define $J = \int_{|\theta|<\alpha} e^{(\rho^2/2)\cdot(e^{i\theta}-1)^2 h''(x,\rho)}d(\rho e^{i\theta})$. The integration path $\gamma = \{|\theta| < \alpha\}$ is part of a circle of radius ρ. We replace it by the path $\gamma_1 \cup \gamma_2 \cup \gamma'_1$ defined as follows: $\gamma_1 = \{y = \rho(1-v) - i\rho\sin\alpha\}$, $\gamma'_1 = \{y = \rho(1-v) + i\rho\sin\alpha\}$ for $v \in [0, 1-\cos\alpha]$, and $\gamma_2 = \{y = \rho + i\rho t\}$, for $t \in [-\sin\alpha, +\sin\alpha]$. See the figure below:



Let J_1, J'_1, and J_2 be the integrals on γ_1, γ'_1, and γ_2: $J = J_1 + J_2 + J'_1$. We first show, below, that the integrals J_1 and J'_1 can be neglected as follows:

$$J_1 = \int_{\gamma_1} e^{(1/2)(y-\rho)^2 h''(x,\rho)}dy$$
$$= -\rho\int_{1-\cos\alpha}^{0} e^{(\rho^2/2)(v+i\sin\alpha)^2 h''(x,\rho)}dv$$
$$= \rho\int_0^{1-\cos\alpha} e^{(\rho^2/2)(v+i\sin\alpha)^2 h''(1,\rho_0)(1+o(1))}dv.$$

Hence $|J_1| \leq \rho\int_0^{1-\cos\alpha} e^{\Re\{\rho^2/2\cdot(v+i\sin\alpha)^2 h''(1,\rho_0)(1+o(1))\}}dv$.

The former integral is $O(\int_0^{1-\cos\alpha} e^{\rho^2/2\cdot(v^2-\sin^2\alpha)h''(1,\rho_0)}dv)$, and

$$\int_0^{1-\cos\alpha} e^{(\rho^2/2)(v^2-\sin^2\alpha)h''(1,\rho_0)}dv = e^{-(\rho^2/2)h''(1,\rho_0)\sin^2\alpha}\int_0^{1-\cos\alpha} e^{(\rho^2/2)v^2 h''(1,\rho_0)}dv.$$

We also have that $\int_0^{1-\cos\alpha} e^{\rho^2/2\cdot v^2 h''(1,\rho_0)}dv \leq (1-\cos\alpha)e^{\rho^2/2\cdot h''(1,\rho_0)(1-\cos\alpha)^2}$. As $h''(1,\rho_0) = \Theta(d)$, this gives for a suitable constant $\gamma_0 > 0$: $J_1 = O(\alpha^2 e^{-\gamma_0 d\alpha^2})$. The

majoration of J_1' is done in the same way. We now show that J_2 gives the main term of J, as follows:

$$\begin{aligned} J_2 &= \int_{\gamma_2} e^{(1/2)(y-\rho)^2 h''(x,\rho)} dy \\ &= i\rho \int_{|t|\le \sin\alpha} e^{-(\rho^2/2)t^2 h''(x,\rho)} dt \\ &= \frac{i}{\sqrt{h''(x,\rho)}} \int_{|v|\le \rho\sqrt{h''(x,\rho)}\sin\alpha} e^{-v^2/2} dv. \end{aligned}$$

This last integral is equal to

$$\int_{-\infty}^{+\infty} e^{-v^2/2} dv - \int_{|v|>\rho\sqrt{h''(x,\rho)}\sin\alpha} e^{-v^2/2} dv = \sqrt{2\pi} + O(e^{-(\rho^2/2)h''(x,\rho)} \sin^2\alpha).$$

This gives $J = i\sqrt{2\pi/h''(x,\rho)}(1 + O(e^{-\gamma_1 d\alpha^2}) + O(\alpha^2\sqrt{d}e^{-\gamma_0 d\alpha^2}))$, with γ_0 and γ_1 strictly positive constants; in particular, they are independent of r and d. We then plug the approximation of J into the expression of the integral to get Lemma B. □

Appendix B. Proof of Lemma C. The main idea in proving Lemma C is a classical one, namely, that the modulus of an analytical function with positive coefficients on a circle of given radius $y > 0$ attains its maximum at point y and only at that point, except when the function is actually a function of y^m for some positive integer m

$$|\lambda(ye^{i\theta})| = |\sum_{n\ge 0} \lambda_n y^n e^{in\theta}| \le \sum_{n\ge 0} \lambda_n y^n = \lambda(y).$$

We want to extend it to get a uniform upper bound $|\lambda(ye^{i\theta})| \le \lambda(y)(1 - C\alpha^2)$, for $\alpha \le |\theta| \le \pi$. We note that, if $\lambda(y) = \Lambda(y^2)$, for example, $|\lambda(ye^{i\theta})|$ would attain its maximum at both points y and $ye^{i\pi}$, and it would not be possible to get an inequality $|\lambda(ye^{i\theta})| \le \lambda(y)(1 - C\alpha^2)$ on most of the circle of radius y, excluding only an arc near y. We first give the proof of Lemma C in a simple case, then the general argument.

Define $\lambda_n = [y^n]\lambda(y)$ for $n \ge 0$. From Property $\mathcal{P}$ of §3.4, we know that $\lambda_0 = 1$. We first assume that $\lambda_1 \ne 0$. We rewrite $\lambda(ye^{i\theta})$ as

$$\lambda(ye^{i\theta}) = (1 + \lambda_1 ye^{i\theta}) + (\lambda(ye^{i\theta}) - 1 - \lambda_1 ye^{i\theta}).$$

The triangular inequality gives

$$|\lambda(ye^{i\theta})| \le |1 + \lambda_1 ye^{i\theta}| + |\lambda(ye^{i\theta}) - 1 - \lambda_1 ye^{i\theta}|.$$

We also have that $\lambda(ye^{i\theta}) - 1 - \lambda_1 ye^{i\theta} = \sum_{n\ge 2} \lambda_n y^n e^{in\theta}$. As the coefficients λ_n are real and positive, we get that

$$|\lambda(ye^{i\theta}) - 1 - \lambda_1 ye^{i\theta}| \le \sum_{n\ge 2} \lambda_n y^n = \lambda(y) - (1 + \lambda_1 y). \tag{22}$$

We can also write

$$|1 + \lambda_1 ye^{i\theta}|^2 = (1 + \lambda_1 y)^2 \left(1 - 2\frac{\lambda_1 y}{(1 + \lambda_1 y)^2}(1 - \cos\theta)\right).$$

This can be simplified by using the fact that $\sqrt{1-t} \leq 1 - t/2$ for $t \in [0,1]$; we get that

$$|1 + \lambda_1 y e^{i\theta}| \leq (1 + \lambda_1 y)\left(1 - \frac{\lambda_1 y}{(1+\lambda_1 y)^2}(1 - \cos\theta)\right).$$

Now, for $|\theta|$ in the interval $[\alpha, \pi]$, we have that $\cos\theta \leq \cos\alpha$; this gives the following inequality valid on $[\alpha, \pi]$, for a strictly positive constant C_0 that can be chosen independent of y and of α (remember that y varies in a compact subset of $]0, +\infty[$):

$$|1 + \lambda_1 y e^{i\theta}| \leq 1 + \lambda_1 y - C_0 \alpha^2. \tag{23}$$

Combined with bound (22), this gives, in turn, the following bound on $|\lambda(ye^{i\theta})|$:

$$|\lambda(ye^{i\theta})| \leq \lambda(y) - C_0\alpha^2 \leq \lambda(y)(1 - C\alpha^2).$$

The term C in this last inequality is positive and can again be chosen so as to be independent of y.

When the term λ_1 is equal to zero, inequality (23) does not hold, and the former proof must be adapted as follows. By Property $\mathcal{P}$, there exists a finite set of indices K such that, for all $k \in K$, $\lambda_k \neq 0$, and that $GCD(k | k \in K) = 1$. Hence there exist relative numbers a_k such that $\sum_k k\, a_k = 1$. Let us define $c = 1/(2\sum_k |a_k|)$; we can assume that α is close enough to zero for the inequality $c\alpha < \pi$ to hold. Let $\theta \in [\alpha, \pi]$ (the case where θ belongs to $[-\pi, -\alpha]$ is symmetrical). Then $\theta = \sum_k a_k k\theta$, and we can show that there is at least one indice $k = k(\theta)$ in K such that $|k\theta[2\pi]| \geq c\alpha$. Assume that it is not the case; then each of the $\eta_k = k\theta[2\pi]$ satisfies $|\eta_k| < c\alpha$; hence $|\theta| = |\sum_k a_k \eta_k| \leq \sum_k |a_k \eta_k| < (\sum_k |a_k|)c\alpha$ and $|\theta| \leq \alpha/2$, which does not hold. We now decompose $\lambda(ye^{i\theta})$ according to the indice $k = k(\theta)$ as follows:

$$\lambda(ye^{i\theta}) = (\lambda(ye^{i\theta}) - 1 - \lambda_k y^k e^{ik\theta}) + (1 + \lambda_k y^k e^{i\theta k}).$$

Hence

$$|\lambda(ye^{i\theta})| \leq \lambda(y) - 1 - \lambda_k y^k + |1 + \lambda_k y^k e^{ik\theta}|.$$

We have that

$$|1 + \lambda_k y^k e^{ik\theta}| \leq (1 + \lambda_k y^k)\left(1 - \frac{\lambda_k y^k}{(1+\lambda_k y^k)^2}(1 - \cos k\theta)\right).$$

As $k\theta[2\pi]$ is at a distance at least $c\alpha$ from 0, we can write

$$|\lambda(ye^{i\theta})| \leq \lambda(y) - \frac{\lambda_k y^k}{1 + \lambda_k y^k}(1 - \cos c\alpha).$$

To get a bound independent of θ, we must remove the dependency of the indice k on θ. This can be done by noting that, as y belongs to a compact set of the reals, the function $a(y) = \min\{\lambda_k y^k : k \in K\}$ is bounded away from zero. This shows the existence of a constant C, independent of θ, such that, for all relevant y and θ, $|\lambda(ye^{i\theta})| \dot{\leq} \lambda(y)(1 - C\alpha^2)$. □

We should point out that, although this is ruled out by our assumptions, there is no difficulty in getting a bound similar to that of Lemma C when function $\lambda(t)$ is affine, of the form $\lambda_0 + \lambda_1 t$, with positive coefficients λ_0 and λ_1. The key condition of our proof is the positivity of the coefficients.

Acknowledgments. The author thanks P. Flajolet for many stimulating discussions on asymptotic distributions, R. Schott for carefully reading a preliminary version of this paper, and an anonymous referee for suggestions that led to a clarification of the original paper, most notably a simpler proof of Lemma C.

REFERENCES

[1] L. AMMANN, *Some limit theorems for clustered occupancy models*, J. Appl. Probab., 20 (1983), pp. 788–802.
[2] E. BENDER, *Central and local limit theorems applied to asymptotic enumeration*, J. Combin. Theory (A), 15 (1973), pp. 91–111.
[3] N. BLEISTEIN AND R. HANDELSMAN, *Asymptotic Expansions of Integrals*, Dover, New York, 1986.
[4] E. R. CANFIELD, *Central and local limit theorems for the coefficients of polynomials of binomial type*, J. Combin. Theory (A), 23 (1977), pp. 275–290.
[5] L. COMTET, *Analyse combinatoire*, Presses Universitaires de France, Paris, 1970.
[6] N. G. DEBRUIJN, *Asymptotic Methods in Analysis*, Dover, New York, 1981.
[7] W. FELLER, *An Introduction to Probability Theory and its Applications*, Vol. 2, John Wiley, New York, 1971.
[8] P. FLAJOLET AND M. SORIA, *Gaussian limiting distributions for the number of components in combinatorial structures*, J. Combin. Theory (A), 53 (1990), pp. 165–182.
[9] D. GARDY, *Bases de données, Allocations aléatoires: Quelques analyses de performances*, Thèse d'Etat, Université de Paris-Sud, Paris, June 1989.
[10] ———, *Join sizes, urn models and normal limiting distributions*, Tech. Report 600, Laboratoire de Recherche en Informatique, Université de Paris-Sud, Paris, October 1990.
[11] D. GARDY AND C. PUECH, *On the sizes of projections: A generating function approach*, Inform. Systems, 9 (1984), pp. 231–235.
[12] ———, *On the effect of join operations on relation sizes*, ACM Trans. Database Systems, 14 (1989), pp. 574–603.
[13] B. GNEDENKO AND A. KOLMOGOROV, *Limit Distributions for Sums of Independent Random Variables*, Addison–Wesley, Reading, MA, 1954.
[14] P. HENRICI, *Applied and Computational Analysis*, Vol. 2, John Wiley, New York, 1977.
[15] M. JARKE AND J. KOCH, *Query optimization in database systems*, ACM Comput. Surveys, 16 (1984), pp. 111–152.
[16] N. JOHNSON AND S. KOTZ, *Urn Models and Their Application*, John Wiley, New York, 1977.
[17] V. KOLCHIN, B. SEVAST'YANOV, AND V. CHISTYAKOV, *Random Allocations*, John Wiley, New York, 1978.
[18] D. MAIER, *The Theory of Relational Databases*, Computer Science Press, New York, 1983.
[19] M. V. MANNINO, P. CHU, AND T. SAGER, *Statistical profile estimation in database systems*, ACM Comput. Surveys, 20 (1988), pp. 191–221.
[20] J. ULLMAN, *Principles of Database Systems*, Computer Science Press, New York, 1980.