ELSEVIER

Theoretical Computer Science 144 (1995) 125–159

Theoretical Computer Science

Dynamic analysis of some relational databases parameters

Danièle Gardy[a,*,1], Guy Louchard[b,2]

[a]*Laboratoire PRISM, Université de Versailles Saint-Quentin, 78035 Versailles, France*
[b]*Département d'Informatique, Université Libre de Bruxelles, Bruxelles, Belgium*

Abstract

We present a dynamic modelization of a relational database, when submitted to a sequence of queries and updates, that allows us to study the evolution of the sizes of relations. These sizes, either present in the database or computed by application of a relational operator (derived relation), have long been recognized as important parameters in query optimization. While the problem of estimating the sizes of derived relations at a given time ("static" case) has been the subject of several studies, to the best of our knowledge the evolution of the relation sizes under queries and updates ("dynamic" case) has not been studied so far.

We consider the size of a relation as a random variable, and we study its probability distribution when the database is submitted to a sequence of insertions, deletions and queries. We show that the relation sizes behave asymptotically as Gaussian processes, whose expectation and covariance are proportional to the time. This approach also allows us to analyse the maximum of the size of the derived relation.

1. Introduction

Among the parameters that can be defined on relational databases, the sizes of the relations, either present in the database or computed by application of a relational operator ("derived" relations) have long been recognized as important parameters in query optimization, i.e. in the search for an efficient way of answering users' queries, and many models have been proposed for their evaluation (see [30] for a survey). So-called *parametric models* are based on a *priori* assumptions on the probability distributions of the objects modelled in the database (relations, attributes, etc); they compute the mean, and sometimes further moments, of the distribution of a derived

[*] Corresponding author.

[1] This research was partly supported by ESPRIT III-Basic Research Action ALCOM II (no. 7141), by the CNRS PRC Mathématique-Informatique and by a cooperation between the CNRS and the FNRS

[2] This research was partially supported by a cooperation between the FNRS and the CNRS.

SSDI 0304-3975(94)00305-X

relation size. Such models are used to estimate the size of a relation obtained by a selection, a projection or a join [2, 16, 17, 35, 36]. *Nonparametric models* use the values present in the database at a given time to obtain empirical information on the underlying probability distributions. This information is summed up in histograms, that are then used to compute estimations of the sizes of derived relations [31, 34], see also [32] for a related approach. An approach popular in recent years is based on *sampling*; again it uses information present in the database to compute estimates of derived relation sizes [18, 19, 24, 33]. All these approaches consider a *static* database, the only exception being the recognized necessity of maintaining some parameters necessary to the sampling process [23].

Our work presents a *parametric model for dynamic databases*: We study the probabilistic behaviour of (initial and derived) relation sizes under assumptions on the values that can be assumed by the database elements, and on the type of operations allowed on the database. As such, it is in close relation to studies on the dynamic behaviour of data structures [9, 11, 22, 26–29].

We gave in former papers [12, 13] conditions which ensure that, in the static case (i.e. at a given time), the size of a derived relation, obtained by a projection, an equijoin or a semijoin, follows a normal limiting distribution. Our goal here is to extend these results to *dynamic databases*, i.e. databases that can be queried and updated. To this effect, we consider the size of a relation as a random variable X, and we study its behaviour when the database is submitted to a sequence of insertions, deletions and queries. We prove that knowing the initial and final sizes of a relation, the constraints on the relation (existence of a functional dependency, sizes of attribute domains, etc.), and the type of operations (queries or updates, with specific probabilities of choosing a given operation at a given time) allows us to characterize completely the random variable X, and that, asymptotically (i.e. for a large number of operations), the size of an initial relation behaves as a Markov Gaussian process, and the size of a derived relation as a (not necessarily Markov) Gaussian process. In both cases, the expectation and covariance are proportional to the time nt, and the process has a deterministic part of order n on which is superimposed a random part of order $\sqrt{n}$. Such a characterization also allows us to analyse the maximum of the size of the derived relation.

The rest of this paper is organized as follows. Section 2 presents the database parameters that we shall study and their modelization in terms of urn models, then briefly recalls the sequences of operations which may be considered. Section 3 gives our main result: the characterization of the size of a derived relation as a Gaussian process, and presents an overview of our method with a sketch of the proof. Section 4 introduces our notations, then Section 5 presents the basic processes (number of tuples in a relation) corresponding to different update models and to several constraints on the initial objects (relations). Sections 6–8 are devoted to the detailed proof of the theorem relative to the projection, Section 9 to the study of the maximum size, and Section 10 to the joins.

2. Databases and urn models

The basic objects we consider are *relations*, which are sets of (distinct) tuples. They can be seen as tables: a row represents a tuple, and the number of lines is the number of elements of the relation (its *size*); the columns are called the *attributes*. The operations we consider on the relations are the projection and the joins (equijoin or semijoin); these relational operations take as arguments one or two relations and define a new relation. For ease of presentation, and without loss of generality, we shall restrict ourselves to the case of relations R or S with two attributes X and Y, or X and Z, and of the projection or the join on X. We shall use the terms *initial relation* for the relations R and S, and *derived relation* for the relation obtained by a projection or a join (see Fig. 1).

We have shown in [12, 13] that it is possible to study the conditional distribution of the sizes of the derived relations, assuming that the sizes of the initial relations are known. To this effect, we introduced, for each operator: projection, equijoin, or semijoin, a modelization in terms of urns and balls that allowed us to see the estimation of the derived relation size as an occupancy model. Now we want to study the variations of this size under a sequence of updates and queries on the database. Again we shall use this modelization, which we briefly recall below.

2.1 Projections and the occupancy problem in urn models

Let d be the number of distinct possible values for the attribute X; we assume that, although it may become large, d is finite. The projection of the relation R can be modelized with urns and balls, according to a well-known *occupancy model*, as follows.

We consider a sequence of d urns, each urn being labelled with a distinct value of the attribute X. To each tuple of the relation R, we associate a ball labelled by the value of tuple on the column X; this ball falls into the corresponding urn. An equivalent way of seeing this phenomenon is to consider instead that we have a finite supply of balls, and

R

X	Y
x_0	y_0
x_0	y_1
x_1	y_2
x_2	y_3

S

X	U
x_0	z_0
x_0	z_1
x_1	z_1
x_3	z_2

$\pi_X(R)$

X
x_0
x_1

$R \bowtie S$

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

$R \rhd S$

X	Y
x_0	y_0
x_0	y_1
x_1	y_2

Fig. 1. Two relations R and S, with the projection $\pi_X(R)$ of R on the attribute X, the equijoin $R \bowtie S$ of R and S and the semijoin $R \rhd S$ of R with S.

that we allocate them at random among the d urns, each trial being independent of the others. Each ball then receives the label of the urn it falls into.

After coupling all the tuples of the initial relation R with urns, some urns are empty and some contain at least one ball. *The number of urns with at least one ball is exactly the number of tuples in the projection of the relation R.*

If, instead of the number of urns with at least one ball, we consider the number of empty urns, and if we assume that each urn can receive an unbounded number of balls, then we have the classical occupancy problem presented for example in [20]. Assuming that the urn size is infinite corresponds, in terms of relational databases, to a relation with a key on the attribute suppressed in the projection. As we shall also want to study relations without keys, we shall have to extend the models to the case where *the urns have a finite capacity* (there are δ places for balls). More generally, if we want to allow for constraints on the database relations, we have to introduce related constraints on the way balls can be allocated into urns (see [12]).

2.2. *Urn models for the equijoin and semijoin*

We have seen that the problem of evaluating the size of the projection of a relation can be reformulated in terms of a classical occupancy problem for a suitable urn model: We throw n balls into a sequence of d distinguishable urns, and study the number of urns with at least one ball. The semijoin and equijoin sizes can likewise be expressed in the general framework of urn models, and we have presented two models to this effect in [13], which we recall below.

Let us start with a sequence of d urns and with two kinds of balls, say blue (B) and red (R); the balls of a given colour are thrown into the urns independently of each other but may depend on the balls of the other colour. After throwing specified numbers of red and blue balls, we assign a certain number of balls of a third colour, say green, to the urns according to one of the two sets of rules below, according to the operation we wish to modelize. The red balls are associated with the relation R, the blue balls with the relation S, and the green balls to their equijoin $R \bowtie S$ or semijoin $R \triangleright S$. The number of balls of one colour is the size of the corresponding relation.

2.2.1. *Model for the equijoin* (*EJ*)

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

2.2.2. *Model for the semijoin* (*SJ*)

- We throw into the urns a given number r of red balls, and a given number s of blue balls.

- For each urn containing at least one blue ball, we put as many green balls as there are red balls. The urns without balls or with balls of only one colour do not receive any green ball.
- We count the total number of green balls.

2.3 Database assumptions

We shall make the following assumptions in the present work, which cover a reasonable number of situations while keeping the computations manageable. We shall assume that *each urn is equally likely*, and that, when the urns have a finite capacity, *each place in an urn is equally likely*. In terms of relational databases, these assumptions mean that the possible values for the projection or join attribute X are uniformly distributed, and that, when the attribute Y or Z, suppressed by the projection or not participating in the join, is not a key of the relation, the possible values of Y or Z are also uniformly distributed. We point out that, when the attribute Y or Z is a key, the (possibly very skewed) probability distribution of the values on this attribute has no influence on the size of the result, as long as we study the distribution of the relation size *conditioned on the initial size* [12, 13].

We also assume that the relations satisfy standard independence assumptions: The coordinates of a tuple are independent, the tuples of a given relation are independent, and, for the join of the relations R and S, the values of the two relations are independent (but see [3, 4] for a discussion about these assumptions).

In the rest of the paper, we shall use indifferently the terms *relation size* and *number of balls* or *number of tuples*, and (in Sections 6–8) the terms *projection size* and *number of nonempty urns*.

2.4. Dynamic models

The urn models we have just defined describe well a relation at a given time, but they do not take into account its evolution during a sequence of updates and queries. We now extend our modelization to consider the evolution of a relation subjected to a sequence of updates (insertions and deletions) and searches (queries).

We denote by $p_{\mathscr{I}}$, $p_{\mathscr{D}}$ and $p_{\mathscr{Q}}$ the probability of making an insertion, a deletion, and a query. If these probabilities vary according to the time t, we use the notations $p_{\mathscr{I}}(t)$, $p_{\mathscr{D}}(t)$ and $p_{\mathscr{Q}}(t)$. We can choose non equal probabilities for insertion and deletion, as long as the probability of an insertion is at least equal to the probability of a deletion: $p_{\mathscr{I}}(t) \geqslant p_{\mathscr{D}}(t)$. Otherwise, the relation is either empty or has very few elements, and this is of little interest, both in terms of database and for the underlying probability model.

We must now make precise the individual probabilities of insertion at a given place, and of deletion of a given ball. If we choose to do a deletion, the conditional probability of deleting a given ball is 1/*number of balls at this time*, both for the infinite urn and for the bounded urn models. If we choose to do an insertion, we must give the conditional probability of inserting a ball into an urn, and the infinite and finite models differ on this point. *In the infinite urn model*, each urn has the same probability of getting the new ball. *If the urns are bounded*, we can view each urn as a collection of δ distinguishable cells, and each empty cell, whatever the urn it belongs to, has the same conditional probability of receiving the ball, given that we have chosen to do an insertion.

To fully specify the dynamic evolution of the relation, we also specify its status at the beginning and at the end of the sequence of updates and queries. We assume that the relation is empty at the beginning. If we impose a condition on the relation at the end (this is not mandatory; see Section 5 for such examples), either the relation is empty or its size is proportional to the time elapsed.

3. Main ideas and results

3.1. The process describing the projection size

Our first goal is to study the variation of the size of the projection under a sequence of queries and updates. We shall do this during a "large" time and for a "large" number of urns. To this effect, we introduce a scaling factor n; the number of urns d is proportional to n and a time τ is written, after normalizations, as $\tau = nt$. The time is chosen in an interval of length $2n$: $0 \leqslant t \leqslant 2$.[3] We shall study two related stochastic processes, describing, respectively, the number of balls, denoted by $\mathscr{P}$, and the size of the projection (number of nonempty urns), denoted by $\mathscr{Q}$; we shall show that each of these processes has a deterministic component of order n, and a random component of order $\sqrt{n}$. Our main result is thus the following theorem, where the functions G, Φ and Ψ_R can be given explicitly for the different models.

Theorem 3.1. *The size $S([nt])$ of the projection at the time nt is asymptotically a (not necessarily Markov) Gaussian process such that*

$$E[S([nt])] \sim nG(t),$$

$$COV(S([nt_1]), S([nt_2])) \sim n\Psi_R(t_1, t_2),$$

$$VAR[S([nt])] \sim n\Phi(t).$$

The relative error in the density due to the asymptotic approximation is $O(1/\sqrt{n})$.

3.2. Sketch of our method

The first step in proving Theorem 3.1 is to study the process $\mathscr{P}$ describing the number of tuples in the initial relation. To fully describe $\mathscr{P}$, we have to know the

[3] We could choose for maximum time n instead of $2n$; however, the second choice gives simpler formulae when the final relation is empty.

probabilities for insertion, deletion and query, and to give the initial and final sizes of the relation. In the cases we are interested in, we can show that $\mathscr{P}$ is a Gaussian processes with a deterministic part $\mathscr{P}_0$, on which is superimposed a random part $\mathscr{P}_1$:

$$\mathscr{P} = \mathscr{P}_0 + \mathscr{P}_1.$$

The process $\mathscr{P}_0$ follows a deterministic curve $nf_1(t)$; the function f_1 is related to the exact relation (or urn) model, and can be computed explicitly. The process $\mathscr{P}_1$ is a Markov Gaussian process of order $\sqrt{n}$. The computation of $\mathscr{P}_0$ and $\mathscr{P}_1$, and of several related parameters such as f_1, is done in Section 5.

The process $\mathscr{P}$ (*number of tuples*) determines another process $\mathscr{Q}$ (*size of the projection*). Before considering $\mathscr{Q}$, we shall study another process $\mathscr{Q}_0$, defined as the size of the projection of a relation R, when the size of R is given by the process $\mathscr{P}_0$ (which is a first-order approximation of $\mathscr{P}$). To this effect, we define two random variables, say Y_1 and Y_2, which are simply the size of the projection at different times t_1 and t_2. We know from previous work [12] that the conditional distribution of the projection size, given the size of the initial relation, follows asymptotically a normal distribution, of known expectation and variance. The covariance $COV(Y_1, Y_2)$ will allow us to characterize $\mathscr{Q}_0$ as a process composed of a deterministic part $nG(t)$ and a random part $\sqrt{n}\,V(t)$. The computation of $COV(Y_1, Y_2)$ starts with Lemma 1 of Section 6.1, and depends on the stochastic behaviour of the number of balls in any one urn. This behaviour can itself be expressed, both for the bounded and for the infinite urn models, in terms of the probabilities $p_{\mathscr{I}}, p_{\mathscr{D}}$ and $p_{\mathscr{Q}}$, and of the function f_1 related to the expectation of the number of balls (Section 6). For any of the processes of Section 5, we could then specialize these results to get the covariance of Y_1 and Y_2; see Sections 7.1 and 7.2 for examples of such computations. We shall rather show that there exists a common form giving the covariance in terms of $f_1, p_{\mathscr{I}}, p_{\mathscr{D}}$ and $p_{\mathscr{Q}}$; this is Proposition 1 of Section 7.3.

We then consider the process $\mathscr{P}$ obtained by superimposing $\mathscr{P}_1$ on $\mathscr{P}_0$. We can again define two random variables *size of the projection* at the times $t_1$and t_2; let us call them S_1 and S_2. As we have done for Y_1 and Y_2, we have to compute their covariance. But the S_i are obtained from the Y_i by introducing a further degree of randomness, and it is possible to write their covariance as

$$COV(S_1, S_2) = COV(Y_1, Y_2) + \gamma(t_1)\gamma(t_2)f_2(t_1, t_2)$$

for a suitable function $\gamma(t), f_2(t_1, t_2)$ being the covariance of the process $\mathscr{P}_1$ taken at different times t_1 and t_2. The covariance of Y_1 and Y_2 thus characterizes the "static" part, and the term added to it to get the covariance of S_1 and S_2 comes from the fact that the number of tuples $\mathscr{P}$ is itself a Gaussian process. The introduction of $\mathscr{P}_1$ and the computation of $COV(S_1, S_2)$ are found in Section 8.

Once we have the covariance of the sizes of the projection at times t_1 and t_2, i.e. of S_1 and S_2, the next part is to show that the final process *size of projection*, which we denote by $\mathcal{Q}$, is still asymptotically a Gaussian process. More precisely, we shall show in Section 8 that $\mathcal{Q}$ has a part $\mathcal{Q}_0$ coming from $\mathcal{P}_0$, on which is added a random part $\mathcal{Q}_1$ coming from $\mathcal{P}_0$ and from $\mathcal{P}_1$:

$$\mathcal{Q} = \mathcal{Q}_0 + \mathcal{Q}_1.$$

3.3. The maximum size of the projection

When we have proved that the final process $\mathcal{Q}$ is Gaussian, and obtained an asymptotic expression for the covariance of S_1 and S_2, we have tools for studying whatever function of the process we are interested in. We shall study here the process giving the maximum size of the projection. We obtain the following result, which is proved in Section 9.

Theorem 3.2. *Let $\bar{t}$ such that $G'(\bar{t}) = 0$. Assume that $\bar{t} \in [0, 2]$ and set $\tilde{G} = G(\bar{t})$. The maximum size of the projection $M := \max_{[0,2]} S([nt])$ occurs at a time t^*, and is such that*

$$M \sim n\tilde{G} + \sqrt{nm} + \mathrm{O}(n^{1/6}),$$

where m and t^* are random variables that can be precisely characterized.

3.4 The process describing the join size

The method we have sketched in Section 3.2 can be adapted to deal with joins. The major modification is that the two initial relations are described by a bi-dimensional process. We obtain the following result, whose proof is given in Section 10.

Theorem 3.3. *In the join model, the size $S([nt])$ of the equijoin or semijoin at the time nt is asymptotically given by a (not necessarily Markov) Gaussian process with*

$$E(S([nt])] \sim nG(t), \quad \textit{with } G(t) := F[f_1^R(t), f_1^B(t)],$$

$$COV(S_1, S_2) \sim n\Psi_R(t_1, t_2),$$

with

$$\begin{aligned}\Psi_R(t_1, t_2) := {} & \Psi_{NR}(t_1, t_2) \\ & + \gamma^R(t_1)\gamma^R(t_2) f_2^{R,R}(t_1, t_2) + \gamma^R(t_1)\gamma^B(t_2) f_2^{R,B}(t_1, t_2) \\ & + \gamma^B(t_1)\gamma^R(t_2) f_2^{B,R}(t_1, t_2) + \gamma^B(t_1)\gamma^B(t_2) f_2^{B,B}(t_1, t_2),\end{aligned}$$

$$\begin{aligned}VAR[S([nt])] \sim {} & n[\Psi_{N,R}(t, t) + \gamma^R(t)^2 f_2^{R,R}(t, t) + 2\gamma^R(t)\gamma^B(t) f_2^{R,B}(t, t) \\ & + \gamma^B(t)^2 f_2^{B,B}(t, t)].\end{aligned}$$

The relative error in the density due to the asymptotic approximation is $\mathrm{O}(1/\sqrt{n})$.

4. Notations

- Let n be some scaling parameter ($n \to +\infty$ later on).
- We have d urns, numbered $i, j, \ldots$; the urn numbered i is U_i. The parameters n and d are related by $d = \Theta(n) = \alpha n$ say.
- When the capacity of an urn is finite, it is denoted by δ (constant). Let $\Delta = d\delta = \beta n$, with $\beta = \alpha\delta$: Δ is the maximal number of balls that we can allocate to the urns.
- Let κ_1^i be the random variable *number of balls in the urn* U_i *at time* t_1, and $E_1^i = E[\kappa_1^i]$; similarly for κ_2^i at time t_2.

  For any measurable function φ and for a random variable κ, we can define a new random variable $\varphi(\kappa)$. Let $E_1^i[\varphi] = E_1^i[\varphi(\kappa_1^i)]$. We shall use in this paper the functions $\varphi(\kappa) = I(\kappa > 0)$ (indicator function) and, in Section 10, $\varphi_2(\kappa) = \kappa$.
- $E_{1,2}^{i,j}$ is the expectation $E[\kappa_1^i \kappa_2^j]$ and similarly $E_{1,2}^{i,j}[\varphi] := E[\varphi(\kappa_1^i)\varphi(\kappa_2^j)]$.
- We denote by n_1 and n_2 the number of tuples of the initial relation, i.e. the total number of balls, at the time t_1 and t_2.
- $\mathscr{I}$, $\mathscr{D}$, $\mathscr{Q}$ denote, respectively, an insertion, a deletion or a query. Their probabilities at the time t are, respectively, $p_{\mathscr{I}}(t)$, $p_{\mathscr{D}}(t)$ and $p_{\mathscr{Q}}(t)$ ($p_{\mathscr{I}}(t) + p_{\mathscr{D}}(t) + p_{\mathscr{Q}}(t) = 1$).
- We denote by $\Rightarrow$ the weak convergence of random functions in the space of all right-continuous functions having left limits and endowed with the Skorohod metric (see [1]). All convergences with be defined for $n \to +\infty$.

5. The process $\mathscr{P}$ related to the initial relation size

Let $W(t)$ be the number of balls at some time t. We might choose the current number of steps (number of queries or updates) as a measure for the time, which would then belong to the interval $[0, 2n]$. However, we shall study the asymptotic behaviour of W when the time goes to infinity, and it is interesting to change the time scale by choosing a time nt for $t \in [0, 2]$, and to normalize the random variable W. For all the models presented below, the number of tuples W has an expectation and a variance of order n, and we can show that, for a suitable function f_1 related to the type of process, and assuming that we start from an empty structure at time 0:

$$\frac{W([nt]) - nf_1(t)}{\sqrt{n}} \Rightarrow X(t), \quad 0 \leqslant t \leqslant 2,$$

where the process $X(t)$ is a Markov Gaussian process whose covariance is denoted $f_2(s, t)$, $s \leqslant t$. As a consequence, we have that for any ξ_1 and ξ_2:

$$E[\mathrm{e}^{\mathrm{i}(\xi_1 n_1 + \xi_2 n_2)}] \sim \exp(n\{\mathrm{i}[\xi_1 f_1(t_1) + \xi_2 f_1(t_2)] \\ - \tfrac{1}{2}[\xi_1^2 f_2(t_1, t_1) + 2\xi_1\xi_2 f_2(t_1, t_2) + \xi_2^2 f_2(t_2, t_2)]\}).$$

We now turn to the presentation of the models we shall study. The processes can be divided in two families:

(i) the *weighted structure* in the sense of Flajolet et al. [10], Louchard [26], with a possibility function given by $pos(\mathscr{D}) = k$ for a k-size structure (there are k ways of deleting an element in a structure composed from k elements!);

(ii) the classical *unweighted structure.*

In the weighted structure family, we have for instance:

- *P*1: $\mathscr{I} + \mathscr{D}$. We a assume that we return to an empty structure at time $2n$. Then [26]

$$f_1(t) = \frac{1}{2}t(2-t), \qquad f_2(s,t) = \frac{s^2}{2}\frac{(2-t)^2}{2},$$

$$p_{\mathscr{I}}(t) = [1 + f_1'(t)]/2 = 1 - t/2, \qquad p_{\mathscr{D}}(t) = [1 - f_1'(t)]/2 = t/2.$$

- *P*2: $\mathscr{I} + \mathscr{D}$. We assume that we return to a structure with size an at time $2n$. The techniques we used in [26] lead here to

$$f_1(t) = t\left(1 - \frac{(2-a)t}{4}\right), \qquad f_2(s,t) = \frac{2-a}{8}s^2(2-t)(-t+at+2), \quad s \leqslant t.$$

  Note that f_1 possesses a maximum for $t \in]0,2[$ iff $a < 1$. For $1 \leqslant a \leqslant 2$, f_1 is maximized at $t = 2$. Again, $p_{\mathscr{I}}(t) = [1 + f_1'(t)]/2$, $p_{\mathscr{D}}(t) = [1 - f_1'(t)]/2$.

- We should be tempted to extend the weighted model to the case with $\mathscr{Q}$. But, when $P(insertion) = P(deletion) = \frac{1}{4}$ and $P(query) = \frac{1}{2}$, we see that the asymptotic total measure along $nf_1(t)$ contains a dominant term $2n \log n C_{\mathscr{D}}(2)$, where $C_{\mathscr{D}}(t)$ is the total number of deletions upto the time t. With constraints on the structure, it leads to $C_{\mathscr{D}}(2) = C_{\mathscr{I}}(2) = n$, i.e. no queries at all, which is a completely uninteresting process!

  So we turn to *the unweighted structure family.*

- *P*3: $\mathscr{I}(p_{\mathscr{I}} \equiv 1)$. We have $n_1 = nt_1$, $n_2 = nt_2$, and $f_1(t) \equiv t$.
- *P*4: *$\mathscr{I} + \mathscr{D} + \mathscr{Q}$ with $p_{\mathscr{I}}$, $p_{\mathscr{D}}$ and $p_{\mathscr{Q}}$ constant ($p_{\mathscr{I}} > p_{\mathscr{D}}$), and without constraint on the relation size at the time* $2n$. The mean and variance corresponding to the variation of the relation size for one step are given by

$$\bar{x} = p_{\mathscr{I}} - p_{\mathscr{D}}, \qquad \sigma^2 = p_{\mathscr{I}} + p_{\mathscr{D}} - \bar{x}^2.$$

  So

$$f_1(t) = \bar{x}t, \qquad f_2(s,t) = \sigma^2 s, \quad s \leqslant t.$$

  This is a classical Brownian motion (BM).

- *P*5: *$\mathscr{I} + \mathscr{D} + \mathscr{Q}$ with arrival at a relation of size $2n\bar{x} + a\sqrt{n}$ at the time* $2n$. The mean $\bar{x}$ and variance σ^2 corresponding to one step are the same as for *P*4, and

$$\frac{W([nt]) - n\bar{x}t}{\sqrt{n}} \Rightarrow \sigma BB(t) + \frac{at}{2},$$

with BB a Brownian bridge. The expectation and covariance of W are given by

$$f_1(t) = \bar{x}t + \frac{at}{2\sqrt{n}}, \qquad f_2(s,t) = \sigma^2 \frac{s(2-t)}{2}, \quad s \leqslant t.$$

- *P6*: $\mathscr{I} + \mathscr{D} + \mathscr{Q}$, *with* $p_{\mathscr{I}}(t)$, $p_{\mathscr{D}}(t)$, $p_{\mathscr{Q}}(t)$ *time dependent*. The infinitesimal mean and variance are given by

  $$\bar{x}(s) = p_{\mathscr{I}}(s) - p_D(s), \qquad \zeta^2(s) = p_{\mathscr{I}}(s) + p_{\mathscr{D}}(s) - \bar{x}^2(s),$$

  so

  $$f_1(t) = \int_0^t \bar{x}(s)\,\mathrm{d}s, \qquad \sigma^2(t) = \int_0^t \zeta^2(s)\,\mathrm{d}s.$$

  The process describing the relation size is a time-dependent BM, which can be written as

  $$BM_0(\sigma^2(t)) = \int_0^t \zeta(s)\,\mathrm{d}BM_1(s),$$

  where BM_0 and BM_1 are standard BM.
- *P7*: *with the* finite urn *case*, assume that each position taken at random among the Δ possible positions changes from status (full → empty, empty → full). This is equivalent to the Ehrenfest urn model ($\mathscr{I} + \mathscr{D}$). From Karlin and Taylor [21, p. 171], we know that, if $\Delta = 2N$, then

  $$\frac{W([Nt]) - N}{\sqrt{N}} \Rightarrow OU(t) \quad (OU(0) = 0 \quad \text{if } W(0) = N).$$

  OU is the classical Ornstein–Uhlenbeck process, with mean 0 and covariance $\frac{1}{2}[\mathrm{e}^{-(t-s)} - \mathrm{e}^{-(t+s)}]$, $s \leqslant t$. Note that this covariance rapidly converges to its stationary form

  $$\tfrac{1}{2}\mathrm{e}^{-(t-s)} \quad \text{as } s, t \to +\infty, \ t - s = \mathrm{O}(1).$$

  Here $\Delta = 2N = \beta n$, so $N = \beta n/2, f_1(t) = \beta/2$ and $p_{\mathscr{I}}(t) = p_{\mathscr{D}}(t) = \frac{1}{2}$. If we start with $W(0) = \beta n/2$, then

  $$\frac{W([nt]) - \frac{\beta}{2}n}{\sqrt{n}} \Rightarrow \sqrt{\frac{\beta}{2}}\, OU\left(\frac{2t}{\beta}\right),$$

  with covariance

  $$f_2(s,t) = \frac{\beta}{4}[\mathrm{e}^{-2(t-s)/\beta} - \mathrm{e}^{-2(t+s)/\beta}].$$

If we start with $W(0)=0$ then it is easily checked that

$$\frac{W([nt])-\frac{\beta}{2}n}{\sqrt{n}}+\frac{\beta}{2}\sqrt{n}\mathrm{e}^{-2t/\beta}\Rightarrow\sqrt{\frac{\beta}{2}}OU\left(\frac{2t}{\beta}\right)$$

and

$$f_1(t)=\frac{\beta}{2}(1-\mathrm{e}^{-2t/\beta}).$$

6. Urn models: preliminary results

6.1. General form of the covariance: Lemma 1

Let Y_1, Y_2 be the sizes of the projection at times t_1, t_2; Lemma 1 below gives a general expression for their covariance $COV(Y_1, Y_2)$, in terms of some probabilities that can be defined whatever the urn model.

We recall that κ_1^i (resp. κ_2^i) is a random variable giving the number of balls in the urn U_i at the time t_1 (resp. t_2). Define $\varphi(\kappa)=I(\kappa>0)$; then the projection sizes Y_1 and Y_2 at the times t_1 and t_2 can be written as $Y_1:=\sum_{i=1}^d\varphi(\kappa_1^i)$ and $Y_2:=\sum_{j=1}^d\varphi(\kappa_2^j)$. Let

$$Z_1^i=Pr[\kappa_1^i=0]=E[I(\kappa_i^1=0)] \tag{1}$$

and similarly for Z_2^i: Z_1^i (resp. Z_2^i) is the probability that the urn U_i is empty at the time t_1 (resp. t_2). Define also the joint probability

$$Z_{1,2}^{i,j}=Pr[\kappa_1^i=0\wedge\kappa_2^j=0]. \tag{2}$$

The term $Z_{1,2}^{i,j}$ is the probability that the urn U_i is empty at the time t_1 *and* that the urn U_j is empty at the time t_2.

Lemma 1. *The expectation of the projection size Y_1 at the time t_1 and the covariance of the projection sizes Y_1 and Y_2 at the times t_1 and t_2 can be expressed in terms of the probabilities Z_1^i, Z_2^j and $Z_{1,2}^{i,j}$ defined by* (1) *and* (2):

$$E(Y_1)=d(1-Z_1^i);\qquad COV(Y_1,Y_2)=d(Z_{1,2}^{i,i}-Z_{1,2}^{i,j})+d^2C_{1,2}^{i,j},$$

with

$$C_{1,2}^{i,j}=Z_{1,2}^{i,j}-Z_1^iZ_2^j.$$

Proof. As the runs are equiprobable:

$$E(Y_1)=\sum_{i=1}^d E[\varphi(\kappa_1^i)]=\sum_{i=1}^d E_1^i[\varphi]=dE_1^i[\varphi].$$

Similarly, $E(Y_2) = dE^j_2[\varphi]$. The probability that the urn U_i is not empty at the time t_1 *is* $E^i_1[\varphi]$, and probability that the urn U_j is not empty at the time t_2 is $E^j_2[\varphi]$:

$$E^i_1[\varphi] = 1 - Z^i_1, \qquad E^j_2[\varphi] = 1 - Z^j_2.$$

By definition, $COV(Y_1, Y_2) = E(Y_1 Y_2) - E(Y_1)E(Y_2)$, and

$$E(Y_1 Y_2) = \sum_{i,j=1}^{d} E[\varphi(\kappa^i_1)\varphi(\kappa^j_2)] = dE^{i,j}_{1,2}[\varphi] + d(d-1)E^{i,j}_{1,2}[\varphi].$$

Now $E^{i,j}_{1,2}[\varphi]$ is the joint probability that the urn U_i is not empty at time t_1, and that the urn U_j is not empty at time t_2. By an argument of inclusion–exclusion, we get

$$E^{i,j}_{1,2}[\varphi] = Pr(\varphi(\kappa^i_1) = \varphi(\kappa^j_2) = 1) = 1 - Z^i_1 - Z^j_2 + Z^{ij}_{1,2}.$$

Hence

$$COV(Y_1, Y_2) = d(Z^i_2 - Z^i_2 + Z^{i,i}_{1,2} - Z^{i,j}_{1,2}) + d^2(Z^{i,j}_{1,2} - Z^i_1 Z^j_2).$$

By symmetry, $Z^j_2 = Z^i_2$, which gives the desired result. □

Let us now turn to the urn models. The stochastic behaviour of a specific urn depends only on $p_{\mathscr{I}}(t)$, $p_{\mathscr{D}}(t)$, $p_{\mathscr{Q}}(t)$ and $f_1(t)$, which in turn depend on the type of process we are concerned with (see Section 5). We shall express our results in terms of the parameters $p_{\mathscr{I}}$, $p_{\mathscr{D}}$, $p_{\mathscr{Q}}$ and f_1, first for infinite urns, then for finite urns. These results can then be specified for any process of Section 5, i.e. for a choice of $p_{\mathscr{I}}$, $p_{\mathscr{D}}$, $p_{\mathscr{Q}}$ and f_1.

6.2. *Model $\mathscr{A}$: the urns are of unlimited size*

We recall that the number of urns d and the parameter n are related by $d = \alpha n$, and that the average number of balls at the time nt is $nf_1(t)$. Following Louchard [27], we see that, asymptotically, the number of balls in a given urn is given by a classical birth and death process with rates

$$\lambda(t) = \frac{p_{\mathscr{I}}(t)}{\alpha}, \qquad \mu(t) = \frac{p_{\mathscr{D}}(t)}{f_1(t)}.$$

The one ball survival probability between t_1 and t_2 is given by

$$ps_{1,2}(t_1, t_2) = ps_{1,2} = \exp\left[-\int_{t_1}^{t_2} \mu(s)\,\mathrm{d}s\right] \ (= 1 \text{ if } t_1 = t_2).$$

The total number of balls inserted in one urn, between t_1 and t_2, and not deleted at t_2, follows a Poisson distribution with parameter

$$\rho_{1,2} = f_3(t_1, t_2)/\alpha, \quad \text{with } f_3(t_1, t_2) := \int_{t_1}^{t_2} p_{\mathscr{I}}(u)\,ps_{1,2}(u, t_2)\,\mathrm{d}u \ (= 0 \text{ if } t_1 = t_2).$$

Note that, by obvious probabilistic reasoning, we have

$$f_1(t_1)ps_{1,2} + f_3(t_1, t_2) = f_1(t_2).$$

6.3. Model $\mathscr{B}$: the urns are bounded

In order to avoid trivial models, we assume that the average number of tuples does not exceed the maximal capacity of all the urns in the time interval we consider: $nf_1(t) \leqslant \Delta$, i.e. $f_1(t) \leqslant \beta$. We first prove two lemmata giving the number of balls in an urn at the time t_1 and its evolution between t_1 and t_2, then analyse the conditional distribution of the number of balls in an urn at the time t_2, conditioned on the number of balls in the urn at the time t_1.

6.3.1. Distribution of the number of balls in an urn

Lemma 2. *At the time t_1, the number of balls κ_1^i in the urn U_i is asymptotically Binomial ($\mathscr{Bin}$) with parameters*

$$\delta,\ f_1(t_1)/\beta.$$

Proof. The probability that there are k balls in the urn U_i is

$$Pr[\kappa_1^i = k] = \frac{\binom{\delta}{\kappa}\binom{\Delta - \delta}{n_1 - k}}{\binom{\Delta}{n_1}} = \binom{\delta}{k}\frac{(\Delta - \delta)!}{\Delta!}\frac{(\Delta - n_1)!}{(\Delta - n_1 - \delta + k)!}\frac{n_1!}{(n_1 - k)!}.$$

Now, for fixed k and δ and large n, and with $\Delta = \beta n$ and $n_1 = nf_1(t_1)$,

$$Pr[\kappa_1^i = k] \sim \binom{\delta}{k}\frac{(\Delta - n_1)^{\delta - k} n_1^k}{\Delta^\delta} \sim \binom{\delta}{k}\left(1 - \frac{f_1(t_1)}{\beta}\right)^{\delta - k}\left(\frac{f_1(t_1)}{\beta}\right)^k,$$

which shows that κ_1^i asymptotically follows a binomial distribution. □

Lemma 3. *Given that we start with k_1 balls in the urn U_i at the time t_1, the number of balls $\kappa(t)$ ($t > t_1$) in the urn U_i is described asymptotically by a birth and death process starting from k_1, with birth rate $\lambda(t) = [\delta - \kappa(t)]f_4(t)$, where $f_4(t) = p_{\mathscr{I}}(t)/[\beta - f_1(t)]$ is the birth rate in a cell, and with individual death rate $f_5(t)$, where $f_5(t) = p_{\mathscr{D}}(t)/f_1(t)$.*

Proof. The probability of insertion in the urn U_i, between t and $t + 1/n$, is

$$p_{\mathscr{I}}(t)\frac{\delta - \kappa(t)}{\Delta - n(t)} = (\delta - \kappa(t))\frac{p_{\mathscr{I}}(t)}{[\beta - f_1(t)]}\frac{1}{n}.$$

Similarly the probability that one tuple is deleted in the urn U_i is

$$p_{\mathscr{D}}(t)\,\frac{\kappa(t)}{n(t)}=\frac{\kappa(t)}{n}\,\frac{p_{\mathscr{D}}(t)}{f_1(t)}. \qquad \square$$

6.3.2. Conditional distribution

To analyse the distribution of the number of balls in an urn at the time t_2, conditioned on the number of balls in the same urn at the time t_1, it is convenient to introduce the function

$$\Pi_{1,2}(k_1,k):=Pr[\kappa(t_2)=k\,|\,\kappa(t_1)=k_1].$$

We can see the content of an urn with $\kappa(t)$ balls as a population of $\kappa(t)$ type 1 (balls) individuals and $\delta-\kappa(t)$ type 2 (empty places) individuals, changing type with rate f_5 and f_4, by Lemma 3. At the time t_1, $\kappa(t_1)=k_1$ and $\delta-\kappa(t_1)=\delta-k_1$. Let

$$p_{i,j}(t_1,t_2)=Pr[\text{individual of type } i \text{ at the time } t_1 \text{ is of type } j \text{ at the time } t_2].$$

So

$$p_{1,2}(t_1,t_2)+p_{1,1}(t_1,t_2)=1, \qquad p_{2,2}(t_1,t_2)+p_{2,1}(t_1,t_2)=1, \tag{3}$$

and the number of balls in the urn at the time t_2 is $\kappa(t_2)=X_{1,1}(t_1,t_2)+X_{2,1}(t_1,t_2)$, where $X_{1,1}(t_1,t_2)$ is the number of balls existing at the time t_1 which are still alive at the time t_2, and $X_{2,1}(t_1,t_2)$ is the number of places empty at the time t_1, and which contain a ball at the time t_2. The random variables $X_{1,1}$ and $X_{2,1}$ are *independent*, with distributions $\mathscr{B}in\,(k_1,p_{1,1}(t_1,t_2))$ and $\mathscr{B}in(\delta-k_1,p_{2,1}(t_1\,t_2))$.

The probability $p_{2,1}(t_1,t_2)$ that a cell which is empty at the time t_1 is full at the time t_2 satisfies the differential equation

$$\begin{aligned}p_{2,1}(t_1,t_2+\mathrm{d}t)={}&p_{2,1}(t_1,t_2)\,Pr[\text{survival during d}t]\\&+p_{2,2}(t_1,t_2)\,Pr[\text{birth during d}t].\end{aligned}$$

Now the probability that there is a birth in an empty cell during an interval of time $\mathrm{d}t$ is $f_4\,\mathrm{d}t$, and the probability that the individual in an occupied cell survives during a time $\mathrm{d}t$ is $1-f_5\,\mathrm{d}t$; hence $p_{2,1}$ satisfies the relation

$$p_{2,1}(t_1,t_2+\mathrm{d}t)=p_{2,1}(t_1,t_2)(1-f_5(t_2)\mathrm{d}t)+p_{2,2}(t_1,t_2)f_4(t_2)\mathrm{d}t.$$

Define

$$f_6(t):=f_4(t)+f_5(t); \qquad f_7(t_1,t_2):=\exp\left[-\int_{t_1}^{t_2}f_6(v)\,\mathrm{d}v\right].$$

With (3), and taking $\mathrm{d}t\to 0$, we readily obtain that the function $t\mapsto p_{2,1}(t_1,t)$ satisfies the differential equation

$$Y'(t)+f_6(t)Y(t)=f_4(t), \quad \text{with } Y(t_1)=0.$$

Solving it, we get

$$p_{2,1}(t_1,t_2)=\int_{t_1}^{t_2} f_4(u)f_7(u,t_2)\mathrm{d}u.$$

The probability $p_{1,1}(t_1,t_2)$ that a cell which is full at the time t_1 is again full at the time t_2 satisfies the same differential equation, but with the initial condition $Y(t_1)=1$; hence

$$p_{1,1}(t_1,t_2)=f_7(t_1,t_2)+p_{2,1}(t_1,t_2)=f_7(t_1,t_2)+\int_{t_1}^{t_2} f_4(u)f_7(u,t_2)\mathrm{d}u.$$

Now the probability $p_{1,2}(t_1,t_2)$ that a cell which is full at the time t_1 becomes empty at the time t_2 satisfies the differential equation

$$Y'(t)+f_6(t)Y(t)=f_5(t), \quad \text{with } Y(t_1)=0,$$

so

$$p_{1,2}(t_1,t_2)=\int_{t_1}^{t_2} f_5(u)f_7(u,t_2)\mathrm{d}u.$$

The probability $p_{2,2}(t_1,t_2)$ that a cell empty at t_1 is still empty at t_2 satisfies the same equation as $p_{1,2}$, but with the initial condition $Y(t_1)=1$; hence

$$p_{2,2}(t_1,t_2)=f_7(t_1,t_2)+p_{1,2}(t_1,t_2)=f_7(t_1,t_2)+\int_{t_1}^{t_2} f_5(u)f_7(u,t_2)\mathrm{d}u.$$

Finally, we obtain the probability $\Pi_{1,2}(k,0)$ that an urn is empty at the time t_2, given that it contains k balls at the time t_1: The k balls become empty places and the $\delta-k$ empty places stay empty; hence

$$\Pi_{1,2}(k,0)=[p_{1,2}(t_1,t_2)]^k\,[p_{2,2}(t_1,t_2)]^{\delta-k}.$$

The conditional expectation of the number of balls at the time t_2, given the number of balls at the time t_1, is

$$M_{1,2}(k)=E[\kappa(t_2)|\kappa(t_1)=k]=kf_7(t_1,t_2)+\delta p_{2,1}(t,t_2).$$

The average number of balls in an urn at the time t is $nf_1(t)/d$; the average numbers of balls in an urn at the times t_1 and t_2, i.e. $nf_1(t_1)/d$ and $nf_1(t_2)/d$, are related by

$$\frac{f_1(t_2)}{d}=\frac{f_1(t_1)}{d}p_{1,1}(t_1,t_2)+\left(\delta-\frac{f_1(t_1)}{d}\right)p_{2,1}(t_1,t_2).$$

Rewriting, we get

$$\frac{f_1(t_1)}{\beta}f_7(t_1,t_2)+p_{2,1}(t_1,t_2)=\frac{f_1(t_2)}{\beta}.$$

7. Computation of $COV(Y_1, Y_2)$ for a nonrandom static structure

In the next subsections, we give some examples of the computation of $COV(Y_1, Y_2)$, before dealing with the general case in Section 7.3. We use the notations $Pi/\mathscr{A}$ for the process Pi in the model with infinite urns, and $Pi/\mathscr{B}$ for the process Pi in the model with bounded urns. The detailed computations can be found in the technical report [14].

7.1. $P3/\mathscr{A}$: *unbounded urns and insertions*

The probability that the urn U_i contains at least one ball at the time t_1 and at the time t_2 is simply the probability that the urn is not empty at the time t_2 (there are no deletions): $Z^{i,i}_{1,2} = Z^i_2$, and Lemma 1 gives

$$COV = d(Z^i_2 - Z^{i,j}_{1,2}) + d^2(Z^{i,j}_{1,2} - Z^i_1 Z^j_2).$$

The next step is to compute the probabilities Z^i_1 and $Z^{i,j}_{1,2}$. We readily derive

$$Z^i_1 = \left(1 - \frac{1}{d}\right)^{n_1} \sim \mathrm{e}^{-t_1/\alpha}, \qquad Z^i_2 = Z^j_2 = \left(1 - \frac{1}{d}\right)^{n_2}.$$

The term $Z^{i,j}_{1,2}$ is the probability that the urn U_i is empty at the time t_1 and that the urn U_j is empty at the time t_2. As there are no deletions, this is also the probability that the urns U_i and U_j are empty at the time t_1 (after throwing n_1 balls) and that the urn U_j is still empty at the time t_2, after throwing $n_{1,2} = n_2 - n_1$ new balls;

$$Z^{i,j}_{1,2} = \left(1 - \frac{2}{d}\right)^{n_1} \left(1 - \frac{1}{d}\right)^{n_{1,2}}.$$

Hence

$$COV = d\left(1 - \frac{1}{d}\right)^{n_{1,2}} \left[\left(1 - \frac{1}{d}\right)^{n_1} - \left(1 - \frac{2}{d}\right)^{n_1}\right]$$
$$+ d^2\left(1 - \frac{1}{d}\right)^{n_{1,2}} \left[\left(1 - \frac{2}{d}\right)^{n_1} - \left(1 - \frac{1}{d}\right)^{2n_1}\right].$$

A direct asymptotic analysis leads to

$$E(Y_1) \underset{n \to \infty}{\sim} \alpha n(1 - \mathrm{e}^{-t_1/\alpha}),$$

$$COV \underset{n \to \infty}{\sim} n\,[\alpha \mathrm{e}^{-t_2/\alpha} - \alpha \mathrm{e}^{-(t_1+t_2)/\alpha} - t_1 \mathrm{e}^{-(t_1+t_2)/\alpha}].$$

However, for further application, it is more convenient to rewrite the covariance in a way that leads to computation of a partial derivative:

$$COV = d\left[\left(1-\frac{1}{d}\right)^{n_1+n_{1,2}} - \left(1-\frac{2}{d}\right)^{n_1}\left(1-\frac{1}{d}\right)^{n_{1,2}}\right]$$

$$+ d^2\left(1-\frac{1}{d}\right)^{n_1+n_{1,2}}\left[\left(1-\frac{1}{d-1}\right)^{n_1} - \left(1-\frac{1}{d}\right)^{n_1}\right], \qquad (4)$$

and we directly analyse the second bracket term of (4). Indeed, this gives an equivalent for $d \sim \alpha n$:

$$\left[\left(1-\frac{1}{d-1}\right)^{n_1} - \left(1-\frac{1}{d}\right)^{n_1}\right] \sim \frac{\partial}{\partial d}\left(1-\frac{1}{d}\right)^{n_1} \sim -\frac{\partial}{\partial \alpha} e^{-t_1/\alpha}\frac{1}{n} = -\frac{t_1}{\alpha^2}\frac{1}{n} e^{-t_1/\alpha},$$

which immediately leads to the asymptotic expression of the covariance.

7.2. $P3/\mathscr{B}$: bounded urns and insertions

In this case, the number of balls at the time t_1 is $n_1 = nt_1$. The approximate value of the probability Z^i_1 that the urn U_i is empty at the time t_1 is given by Lemma 2, with $f_1(t) = t$:

$$Z^i_1 \sim \left(1-\frac{t_1}{\beta}\right)^{\delta}.$$

The probability that the urn U_i is empty at the time t_1 and at the time t_2 is again simply the probability that it is empty at the time t_2:

$$Z^{i,i}_{1,2} \sim \left(1-\frac{t_2}{\beta}\right)^{\delta}.$$

The joint probability that the urn U_i is empty at the time t_1 and that the urn U_j is empty at the time t_2 is

$$Z^{i,j}_{1,2} \sim \left(1-\frac{t_1}{\beta}\right)^{\delta}\left(1-\frac{t_1}{\beta'}\right)^{\delta}\left(1-\frac{t_2-t_1}{\beta''}\right)^{\delta},$$

with $\beta' = \beta - \delta/n$ and $\beta'' = \beta - t_1$. So, asymptotically, $Z^{i,j}_{1,2} \sim (1-t_1/\beta)^{\delta}(1-t_2/\beta)^{\delta}$. By $C^{i,j}_{1,2} = Z^{i,j}_{1,2} - Z^i_1 Z^j_2$, we obtain

$$C^{i,j}_{1,2} \sim \left(1-\frac{t_1}{\beta}\right)^{\delta}\left[\left(1-\frac{t_1}{\beta'}\right)^{\delta} - \left(1-\frac{t_1}{\beta}\right)^{\delta}\right]\left(1-\frac{t_2-t_1}{\beta''}\right)^{\delta}.$$

After some computations, and with Lemma 1, we obtain

$$E(Y_1) \sim \alpha n\left[1-\left(1-\frac{t_1}{\beta}\right)^{\delta}\right],$$

$$COV \sim n\left[\alpha\left(1-\frac{t_2}{\beta}\right)^{\delta}-\alpha\left(1-\frac{t_1}{\beta}\right)^{\delta}\left(1-\frac{t_2}{\beta}\right)^{\delta}-t_1\left(1-\frac{t_1}{\beta}\right)^{\delta-1}\left(1-\frac{t_2}{\beta}\right)^{\delta}\right].$$

7.3. *Nonrandom static structure*

We now extend the computations of Sections 7.1 and 7.2 to the general cases. Let us start with a nonrandom (NR) static structure, where the process *total number of tuples* follows a path fixed by $nf_1(t)$ i.e. the path is the closest to $nf_1(t)$ in the adequate topology (Skorohod for instance). On a short interval $Dt \ll 1$, such that $n\,Dt \gg 1$, each step chosen at random among m has probability $p_{\mathscr{I}}(t)$, $p_{\mathscr{D}}(t)$ or $p_{\mathscr{Q}}(t)$ of giving an insertion, a deletion or a query.

All the results in this section depend only on the function $f_1(t)$ specifying the number of balls at time t, on the probabilities $p_{\mathscr{I}}(t)$, $p_{\mathscr{D}}(t)$ and $p_{\mathscr{Q}}(t)$, and on the auxiliary functions defined in Sections 6.2 and 6.3. We shall show that the average value of the projection size at time t_1, $E(Y_1)$, and its covariance at distinct times t_1 and t_2, $COV(Y_1, Y_2)$, both have a common form whatever the model:

Proposition 1. *For each urn model $\mathscr{A}$ or $\mathscr{B}$, there exist two functions $F(x)$ and $\Psi_{NR}(t_1, t_2)$ such that, if we consider a relation of size $nf_1(t)$, the size of its projection is a random variable with expectation at the time t_1 $E(Y_1)$, and covariance at distinct times t_1 and t_2 $COV(Y_1, Y_2)$, given asymptotically by:*

$$E(Y_1) \sim nF(f_1(t_1)), \qquad COV(Y_1, Y_2) \sim n\Psi_{NR}(t_1, t_2).$$

We shall study two examples before proving Proposition 1 in Section 7.4.3.

7.3.1. *Unbounded urns and return to an empty structure: $P1/\mathscr{A}$*

We use the notations $Z_1^i(\mathscr{A})$ and so on to emphasize the dependence on the urn model. We obtain

$$E(Y_1) = \alpha n(1 - Z_1^i(\mathscr{A})) \sim \alpha n(1 - \mathrm{e}^{-f_1(t_1)/\alpha}) = nF(f_1(t_1)) \text{ say,}$$

with $f_1(t) = t(2-t)/2$. By Lemma 1,

$$COV(Y_1, Y_2) = \alpha n[Z_{1,2}^{i,i}(\mathscr{A}) - Z_{1,2}^{i,j}(\mathscr{A})] + \alpha^2 n^2 C_{1,2}^{i,j}(\mathscr{A}).$$

So we obtain

$$\begin{aligned} COV \underset{n\to+\infty}{\sim} & n[\alpha \mathrm{e}^{-[f_1(t_1)+f_3(t_1,t_2)]/\alpha} - \alpha \mathrm{e}^{-[f_1(t_1)+f_1(t_2)]/\alpha} \\ & - ps_{1,2} f_1(t_1) \mathrm{e}^{-[f_1(t_1)+f_1(t_2)]/\alpha}] \\ = {} & n\Psi_{NR}(t_1, t_2) \end{aligned}$$

and $ps_{1,2}\, f_1(t_1) = t_1(2-t_2)/2, f_3(t_1, t_2) = (1 - t_2/t_1)(t_2 - t_1)$.

7.3.2. *Bounded urn and queries: P4/$\mathscr{B}$*

We obtain

$$E(Y_1) =: nF(f_1(t_1)) \sim \alpha n\left(1-\left(1-\frac{f_1(t_1)}{\beta}\right)^{\delta}\right),$$

$$\begin{aligned}COV(Y_1, Y_2) &=: n\Psi_{NR}(t_1, t_2)\\ &\sim \alpha n\left(1-\frac{\bar{x}t_1}{\beta}\right)^{\delta}\left[p_{2,2}(t_1,t_2)^{\delta}-\left(1-\frac{\bar{x}t_2}{\beta}\right)^{\delta}\right]\\ &\quad - n\bar{x}t_1 f_7(t_1,t_2)\left(1-\frac{\bar{x}t_1}{\beta}\right)^{\delta}\left(1-\frac{\bar{x}t_2}{\beta}\right)^{\delta-1}.\end{aligned}$$

7.3.3 *Proof of Proposition* 1

From Lemma of Section 6.1, we know that $E(Y_1) = d(1 - Z_1^i) = n\alpha(1 - Z_1^i)$. After some computations (given in [14]), we get

- For the model with unbounded urns, $Z_1^i \sim \mathrm{e}^{-f_1(t)/\alpha}$, and $E(Y_1) \sim nF(f_1(t))$ with $F(x) = \alpha(1 - \mathrm{e}^{-x/\alpha})$. The covariance is

$$\begin{aligned}COV(Y_1, Y_2) \sim n(&\alpha\mathrm{e}^{-(f_1(t_1)+f_3(t_1,t_2))/\alpha} - \alpha\mathrm{e}^{-(f_1(t_1)+f_1(t_2))/\alpha}\\ &-f_1(t_1)ps_{1,2}\mathrm{e}^{(f_1(t_1)+f_1(t_2))/\alpha}),\end{aligned}$$

  hence $COV(Y_1, Y_2) \sim n\Psi_{NR}(t_1, t_2)$ for a function $\Psi_{NR}(t_1, t_2)$ which is expressed in terms of the function $f_1(t)$ and $f_3(t_1, t_2)$:

$$\Psi_{NR}(t_1,t_2) = \alpha\mathrm{e}^{-f_1(t_1)/\alpha}(\mathrm{e}^{-f_3(t_1,t_2)/\alpha} - \mathrm{e}^{-f_1(t_2)/\alpha}) - f_1(t_1)ps_{1,2}\mathrm{e}^{-(f_1(t_1)+f_1(t_2))/\alpha}.$$

- For the model with bounded urns, $Z_1^i \sim (1 - f_1(t)/\beta)^{\delta}$, and $E(Y_1) \sim nF(f_1(t))$ with $F(x) = \alpha(1 - (1 - x/\beta)^{\delta})$. Using the functions defined in Section 6.3, we get an expression of the covariance as

$$\begin{aligned}COV(Y_1, Y_2) \sim n\Bigg(&\alpha\left(1-\frac{f_1(t_1)}{\beta}\right)^{\delta}\left[p_{2,2}(t_1,t_2)^{\delta}-\left(1-\frac{f_1(t_2)}{\beta}\right)^{\delta}\right]\\ &-f_1(t_1)f_7(t_1,t_2)\left(1-\frac{f_1(t_1)}{\beta}\right)^{\delta}\left(1-\frac{f_1(t_2)}{\beta}\right)^{\delta-1}\Bigg),\end{aligned}$$

  hence $COV(Y_1, Y_2) \sim n\Psi_{NR}(t_1, t_2)$ for a suitable function $\Psi_{NR}(t_1, t_2)$ given below (we recall that $p_{2,2}$ and f_7 are functions of t_1 and t_2):

$$\begin{aligned}&\Psi_{NR}(t_1,t_2)\\ &= \left(1-\frac{f_1(t_1)}{\beta}\right)^{\delta}\left[\alpha p_{2,2}^{\delta}-\left(1-\frac{f_1(t_2)}{\beta}\right)^{\delta-1}\left(\alpha p_{2,2}+\left(1-\frac{1}{\delta}\right)f_1(t_1)f_7\right)\right].\end{aligned}$$

8. Random structure

We can generalize the techniques we used in [27]. We recall that Y_1 and Y_2 denote the size of the projection of a relation R at the times t_1 and t_2, when the number of tuples of R is given by the process $\mathscr{P}_0$, and that S_1 and S_2 denote the same quantities when the number of tuples of R is given by the process $\mathscr{P}$. We first compute the variation of $COV(Y_1, Y_2)$ introduced by assuming that the numbers of tuples are no longer fixed, but Gaussian random variables; this gives a term that we call $n\Psi_C(t_1, t_2)$. Then we compute the actual covariance of S_1 and S_2 and we show that it is of the type $n\Psi_R(t_1, t_2)$; its form shows that the size of the projection is a Gaussian process. Below we state our result, before proving it in the rest of this section. In the following theorem, f_1 and f_2 are relative to the expectation and covariance of the process associated with the initial relation, and $\gamma(t) = F'(f_1(t))$.

Theorem 8.1. *In the projection model, the size of the projection* $S([nt])$ *is a (not necessarily Markov) Gaussian process with*

$$E[S([nt])] \sim nG(t), \quad \textit{with } G(t) := F(f_1(t)),$$

$$COV(S([nt_1]), S([nt_2])) \sim n\Psi_R(t_1, t_2),$$

$$\textit{with } \Psi_R(t_1\ t_2) := \Psi_{NR}(t_1, t_2) + f_2(t_1, t_2)\gamma(t_1)\gamma(t_2),$$

$$VAR[S([nt])] \sim n\Phi(t) \quad \textit{with } \Phi(t) := \Psi_R(t, t) = \Psi_{NR}(t, t) + \gamma^2(t) f_2(t, t).$$

The relative error in the density is $O(1/\sqrt{n})$.

8.1. The perturbation on the expectation and covariance of Y_1 and Y_2

In this part, we take into account the random part of the number of balls $W([nt])$. As the process $\mathscr{P}$ is obtained by adding a process $\mathscr{P}_1$ or order $\sqrt{n}$ to the process $\mathscr{P}_0$, itself of order n, the respective numbers of balls at the times t_1 and t_2 are

$$n_1 = n\left(f_1(t_1) + \frac{\theta_1}{\sqrt{n}}\right) + O(1), \qquad n_2 = n\left(f_1(t_2) + \frac{\theta_2}{\sqrt{n}}\right) + O(1), \tag{5}$$

where θ_1 and θ_2 are Gaussian random variables with mean 0 and covariance $f_2(s, t)$: for any ζ_1 and ζ_2,

$$E[e^{i(\zeta_1\theta_1 + \zeta_2\theta_2)}] \sim e^{-1/2[f_2(t_1, t_1)\zeta_1^2 + 2f_2(t_1, t_2)\zeta_1\zeta_2 + f_2(t_2, t_2)\zeta_2^2]}. \tag{6}$$

Set

$$\gamma(t) := F'(f_1(t)). \tag{7}$$

From Proposition 1 of Section 7, we have that $E[Y_1] \sim nF(n_1/n)$, and with (5) this gives

$$E[Y_1] \sim n\left(F(f_1(t_1)) + \frac{\theta_1}{\sqrt{n}}\gamma(t_1)\right).$$

A similar relation holds for $E[Y_2]$. Now, injecting the values of n_1 and n_2 given by the formulae (5) into the covariance $COV(Y_1, Y_2) = n\Psi_{NR}(t_1, t_2)$, we get a new value $n\Psi_C(t_1, t_2)$, with

$$\Psi_C(t_1, t_2) = \Psi_{NR}(t_1, t_2) + \bar{\varphi}_1(t_1, t_2)\frac{\theta_1}{\sqrt{n}} + \bar{\varphi}_2(t_1, t_2)\frac{\theta_2}{\sqrt{n}} + \mathrm{O}\left(\frac{1}{n}\right)$$

for some $\bar{\varphi}_1$ and $\bar{\varphi}_2$, and for Ψ_{NR} computed in Section 7.3.3 according to the type of urn.

8.2. *The covariance of* S_1 and S_2

We know from previous work [12] that, for a known size of the initial relation R at the times t_1 and t_2, the projection sizes Y_1 and Y_2 are asymptotically normal. Then for any ξ_1 and ξ_2

$$E[\mathrm{e}^{\mathrm{i}(\xi_1 Y_1 + \xi_2 Y_2)}] \sim E[\mathrm{e}^{\mathrm{i}(\xi_1 E[Y_1] + \xi_2 E[Y_2]) - 1/2(\xi_1^2\sigma^2(Y_1) + 2\xi_1\xi_2 COV(Y_1, Y_2) + \xi_2^2\sigma^2(Y_2))}].$$

Plugging into this equation the modified values for $E[Y_1]$ and $E[Y_2]$, and substituting $n\Psi_C(t_1, t_2)$ for $COV(Y_1, Y_2)$ (and similarly for $\sigma^2(Y_1)$ and $\sigma^2(Y_2)$), we obtain an expression for the expectation $E[\mathrm{e}^{\mathrm{i}(\xi_1 S_1 + \xi_2 S_2)}]$:

$$E[\mathrm{e}^{\mathrm{i}(\xi_1 S_1 + \xi_2 S_2)}] \sim \mathrm{e}^{A(t_1, t_2)} E[\mathrm{e}^{B(t_1, t_2)}],$$

with

$$A(t_1, t_2) = \mathrm{i}[\xi_1 nF(f_1(t_1)) + \xi_2 nF(f_1(t_2))]$$
$$\qquad - \frac{1}{2}n[\Psi_{NR}(t_1, t_1)\xi_1^2 + 2\Psi_{NR}(t_1, t_2)\xi_1\xi_2 + \Psi_{NR}(t_2, t_2)\xi_2^2]$$

and

$$B(t_1, t_2) = \mathrm{i}\theta_1\sqrt{n}\left(\xi_1\gamma(t_1) + \frac{\mathrm{i}}{2}\bar{\varphi}_1(t_1, t_1)\xi_1^2 + \mathrm{i}\bar{\varphi}_1(t_1, t_2)\xi_1\xi_2 + \frac{\mathrm{i}}{2}\bar{\varphi}_1(t_2, t_2)\xi_2^2\right)$$
$$+ \mathrm{i}\theta_2\sqrt{n}\left(\xi_2\gamma(t_2) + \frac{\mathrm{i}}{2}\bar{\varphi}_2(t_2, t_2)\xi_2^2 + \mathrm{i}\bar{\varphi}_2(t_1, t_2)\xi_1\xi_2 + \frac{\mathrm{i}}{2}\bar{\varphi}_2(t_1, t_1)\xi_1^2\right) + \mathrm{O}(1).$$

The term $B(t_1, t_2)$ contains all the contribution from the random variables θ_1 and θ_2 and is of the form $\mathrm{i}(\zeta_1\theta_1 + \zeta\theta_2)$; this leads, with (6), to

$$E[\mathrm{e}^{B(t_1,t_2)}] \sim \exp(-\tfrac{1}{2}n\,[\xi_1^2\gamma^2(t_1)f_2(t_1,t_1) + 2\xi_1\xi_2\gamma(t_1)\gamma(t_2)f_2(t_1,t_2)$$
$$+ \xi_2^2\gamma^2(t_2)f_2(t_2,t_2) + \text{cubic terms in } \xi_1, \xi_2 + \mathrm{O}(1/\sqrt{n})]).$$

Let us define

$$G(t) = F(f_1(t)), \qquad \Psi_R(t_1,t_2) = \Psi_{NR}(t_1,t_2) + f_2(t_1,t_2)\gamma(t_1)\gamma(t_2).$$

We get

$$E[\mathrm{e}^{\mathrm{i}(\xi_1S_1 + \xi_2S_2)}] \sim \exp(\mathrm{i}\,[\xi_1nG(t_1) + \xi_2nG(t_2)]$$
$$-\tfrac{1}{2}[\xi_1^2n\Psi_R(t_1,t_1) + 2\xi_1\xi_2n\Psi_R(t_1,t_2) + \xi_2^2n\Psi_R(t_2,t_2)]$$
$$+ \text{cubic terms in } \xi_1, \xi_2 + \mathrm{O}(1/\sqrt{n})).$$

Now we remember that we are actually interested in the normalized process $S'([nt]) = (S([nt]) - nG(t))/\sqrt{n}$. Substituting $\xi'_1/\sqrt{n}$ for ξ_1 and $\xi'_2/\sqrt{n}$ for ξ_2, we get

$$E[\mathrm{e}^{\mathrm{i}(\xi'_1S'_1 + \xi'_2S'_2)}] \sim \exp(-\tfrac{1}{2}\xi_1'^2\Psi_R(t_1,t_1) + 2\xi'_1\xi'_2\Psi_R(t_1,t_2) + \xi_2'^2\Psi_R(t_2,t_2)]$$
$$+ \mathrm{O}(1/\sqrt{n})).$$

which proves Theorem 8.1.

8.3. Examples

Let us illustrate our technique with two examples. We choose two distinct processes, corresponding to two types of operation in the infinite model. In Section 7.3, we derived the following results for the model with unbounded urns: $\gamma(t) = \mathrm{e}^{-f_1(t)/\alpha}$ (see (7)) and

$$G(t) = F(f_1(t)) = \alpha(1 - \mathrm{e}^{-f_1(t)/\alpha});$$

$$\Psi_{N,R}(t_1,t_2) = \alpha\mathrm{e}^{-[f_1(t_1)+f_3(t_1,t_2)]/\alpha} - \alpha\mathrm{e}^{-[f_1(t_1)+f_1(t_2)]/\alpha}$$
$$- ps_{1,2}f_1(t_1)\mathrm{e}^{-[f_1(t_1)+f_1(t_2)]/\alpha}.$$

1. *P1/𝒜: weighted process with insertions and deletions.* For a sequence of insertions and deletions, starting from and arriving at an empty relation, we have $f_1(t) = t(2-t)/2$ and $f_2(t_1,t_2) = t_1^2(2-t_2)^2/4$. The survival probability is $ps_{1,2}(t_1,t_2) = (2-t_2)/(2-t_1)$, and $f_3(t_1,t_2) = (1-t_2/2)(t_2-t_1)$. Theorem 8.1 leads to

$$E[S([nt])] \sim n\alpha(1 - \mathrm{e}^{-t(2-t)/2\alpha}),$$

$$COV(S_1,S_2) \sim n\mathrm{e}^{-\sigma/\alpha}[\alpha(\mathrm{e}^{\tau/\alpha} - 1) + \tau^2] - \tau\mathrm{e}^{\sigma/\alpha},$$

with

$$\tau = \frac{t_1(2-t_2)}{2} \quad \text{and} \quad \sigma = \frac{t_1(2-t_1)}{2} + \frac{t_2(2-t_2)}{2}.$$

2. *$P5/\mathscr{A}$: unweighted process with insertions, deletions and queries.* We study now a sequence of insertions, deletions and queries, starting from an empty structure and arriving at a structure of size $2n\bar{x} + a\sqrt{n}$. We have $F(x)$, $\gamma(t)$ and Ψ_{NR} as above, but the functions f_1 and f_2 are those corresponding to $P5$:

$$f_1(t) = \bar{x}t + \frac{at}{2\sqrt{n}}, \qquad f_2(s,t) = \sigma^2 \frac{s(2-t)}{2},$$

(we recall that $\bar{x} = p_{\mathscr{I}} - p_{\mathscr{D}}$), and $ps_{1,2}$ and f_3 are as follows:

$$ps_{1,2}(t_1,t_2) = (t_1/t_2)^{p_{\mathscr{D}}/\bar{x}}, \qquad f_3(t_1,t_2) = \bar{x}(t_2 - t_1(t_1/t_2)^{p_{\mathscr{D}}/\bar{x}}).$$

We give below the expectation and the covariance of the process *size of the projection* in this case:

$$E[S([nt])] \sim n\alpha(1 - \mathrm{e}^{-\bar{x}t/\alpha}) + \sqrt{n}\, ate^{-\bar{x}t/\alpha},$$

$$COV(S_1,S_2) \sim n\Psi_R(t_1,t_2),$$

with

$$\Psi_R(t_1,t_2) = \mathrm{e}^{-\bar{x}(t_1+t_2)/\alpha}\left[\alpha(\mathrm{e}^{-\bar{x}(t_1(t_1/t_2)^{p_{\mathscr{D}}/\bar{x}}/\alpha} - 1) - \bar{x}t_1\left(\frac{t_1}{t_2}\right)^{p_{\mathscr{D}}/\bar{x}} + \sigma^2\frac{t_1(2-t_2)}{2}\right].$$

9. Projection maximum

We have shown in Section 8 that the projection size at the time nt is

$$S([nt]) = nG(t) + \sqrt{n}V(t) + \mathrm{O}(1),$$

with G a deterministic curve and V a Gaussian process. To analyse the maximum of $S([nt])$, we must first know whether $G(t)$ has a maximum for $t \in]0,2[$. If $G(t)$ is maximized at $t = 2$ then it is easily checked that $S([nt])$ is also maximized at $t = 2$. So assume that $G'(\bar{t}) = 0$ for a unique $\bar{t} \in]0,2[$ (local maxima can be analyzed similarly). Set $\bar{G} := G(\bar{t})$ and $\bar{G}'' = G''(\bar{t})$.

It is equivalent to analyse the maximum of S, or the maximum of the normalized process obtained by a change of scale in the process and in the time:

$$X(t) := \frac{S([nt]) - n\bar{G}}{\sqrt{n}} \sim V(t) + \sqrt{n}(G(t) - \bar{G}). \tag{8}$$

We shall use a technique based on the results of Daniels [6], which applies precisely to a process of the form (8), and which we recall below.

9.1. The basic results

Consider a Gaussian process $V(t)$ superimposed on a curve $\tilde{y}(t)$. It is equivalent to look for its maximum $m := max\,[V(t) + \tilde{y}(t)]$, and the time t^* at which this maximum occurs, or to search for the hitting time of $V(t)$ to the absorbing boundary $m - \tilde{y}(t)$. It is well known (see [8]) that, near the crossing point, $V(t)$ behaves locally like a Brownian motion BM, or a variant of it, such as a Brownian bridge BB. It is also known that the hitting time and place densities for a BB can be deduced from the hitting time density for a BM (see for instance Louchard [25] for a constant boundary and Csaki et al. [5] for a general proof).

Assume that $\tilde{y}(t)$ is given by

$$\tilde{y}(t) = \sqrt{n}\, y(t), \quad n \gg 1 \tag{9}$$

and that it has a unique maximum at $\bar{t}$, with $y(\bar{t}) = 0$ (if necessary, we translate the origin). Daniels and Skyrme [7] have computed the asymptotic hitting time and place density. In the Gaussian process case, with covariance $C(s, t)$, $s \leqslant t$, Daniels [6] has matched the local behaviour of $C(s, t)$ with the BM (or one of its variants) covariance near $\bar{t}$. In the BB match, we have

$$[V(t) + \sqrt{n}y(t)] \sim \sqrt{A}[BB(t - t_0) + \sqrt{n}\varphi(t - t_0)] \quad \text{on } t \in (t_0, t_0 + T),$$

where $BB(T) = 0$, $y(t) \equiv \varphi(t - t_0)$ and A is some constant. We can deduce the density of the maximum m and time t^* from Daniels [6, (3.8)] and Daniels and Skyrme [7, (5.9)]. We first need to introduce some notations. Let

$$c_1 := [\partial_s C(s, t)]_{\bar{t}} \geqslant 0, \qquad c_2 := [\partial_t C(s, t)]_{\bar{t}} \leqslant 0, \qquad c := C(\bar{t}, \bar{t}), \tag{10}$$

$$t_0 := \bar{t} - c/c_1, \qquad t_0 + T := \bar{t} + c/|c_2|, \qquad T := cA/(c_1|c_2|), \tag{11}$$

$$A := c_1 + |c_2|, \qquad B := -y''(t), \qquad u := n^{1/3}A^{-1/3}B^{2/3}(t^* - \bar{t}). \tag{12}$$

Let also $R(x) := \exp(x^3/6)H(x)$, with

$$H(x) := 2^{-1/3}\frac{1}{2\pi \mathrm{i}}\int_{-\mathrm{i}\infty}^{+\mathrm{i}\infty} \mathrm{e}^{sx}\frac{\mathrm{d}s}{A_i(-2^{1/3}s)},$$

A_i is the classical Airy function. Let $f(x) := 2R(x)R(-x)$ and $v(x) := H'(x)/H(x)$; R and v are tabuled in Daniels [6]; note that $f'(0) = 0$. Finally, define

$$\lambda := \int_{-\infty}^{+\infty} [R(x) - x^+]\,\mathrm{d}x = 0.99615\ldots$$

The result of Daniels and Skyrme is as follows:

Theorem 9.1. *The random variable* $m := max[V(t) + \tilde{y}(t)]$ *is asymptotically Gaussian with mean and variance*

$$E(m) \sim \lambda n^{-1/6} A^{2/3} B^{-1/3}, \qquad \sigma^2(m) \sim c. \tag{13}$$

The conditioned maximum $m \,|\, t^*$ *is asymptotically Gaussian with mean and variance*

$$E(m \,|\, t^*) \sim n^{-1/6} A^{-1/3} [c_1 v(-u) + |c_2| v(u)] B^{1/3}, \qquad \sigma^2[m \,|\, t^*] \sim c. \tag{14}$$

The joint density of m *and* t^* *is given by*

$$\varphi(m, u) \mathrm{d}m \, \mathrm{d}u = 2 \sqrt{\frac{1}{2\pi c}} \mathrm{e}^{-m^2/(2c)} \left\{ \frac{f(u)}{2} + n^{-1/6} B^{-1/3} A^{-1/3} m \varphi_1(m, u) \right.$$
$$\left. + \, \mathrm{O}(n^{-1/3}) \right\} \mathrm{d}m \, \mathrm{d}u, \tag{15}$$

with

$$\varphi_1(m, u) = -\frac{1}{4} u^2 f(u) \frac{A}{c} + R'(-u) R(u) \frac{c_1}{c} + R(-u) R'(u) \frac{|c_2|}{c}, \tag{16}$$

and where u *has density* $f(u)$. *All expectations and densities have relative errors of order* $\mathrm{O}(n^{-1/3})$.

9.2. Application to the projection size

From Theorem 8.1, we know that

$$S([nt]) = \sqrt{n} \{ \sqrt{n} \bar{G} + V(t) + \sqrt{n} [G(t) - \bar{G}] \} + \mathrm{O}(1),$$

with $COV[V(t_1), V(t_2)] = \Psi_R(t_1, t_2)$. The O(1) term is non-uniform in V but it is easy to check that $max(S)$ is only affected by a O(1) term. Comparing with (9), we must identify $y(t)$ with $G(t) - \bar{G}$. We can now compute

$$c_1 := \partial_{t_1} \Psi_R(t_1, t_2)|_{t_1 = t_2 = \bar{t}}, \qquad c_2 := \partial_{t_2} \Psi_R(t_1, t_2)|_{t_1 = t_2 = \bar{t}}, \tag{17}$$

$$c_1 := \Psi_R(\bar{t}, \bar{t}), \qquad A := c_1 + |c_2|, \qquad B := -\bar{G}''. \tag{18}$$

The result of Daniels leads to the following theorem.

Theorem 9.2. *The maximum size of the projection satisfies*

$$M := \max_{[0,2]} S([nt]) \sim n\tilde{G} + m\sqrt{n} + \mathrm{O}(n^{1/6}),$$

where m *is a random variable characterized by* (10)–(15). *The time* t^* *when the maximum occurs is a random variable characterized by* (12), (15) *and* (16). *The constants are given by* (17) *and* (18).

Let us again illustrate our theorem with an example. For the unbounded urn model, and the process P_1 (no deletions or queries), the function $G(t)$ is maximized for $\bar{t} = 1$ so $\bar{G} = \alpha(1 - \mathrm{e}^{-1/2\alpha})$ and

$$B = -\bar{G}'' = \mathrm{e}^{-1/2\alpha}, \qquad c = \alpha^{-1/2\alpha} - \alpha e^{-1/\alpha} - \tfrac{1}{4}\mathrm{e}^{-1/\alpha},$$

$$c_1 = \tfrac{1}{2}\mathrm{e}^{-1/2\alpha}, \quad c_2 = -c_1, \qquad A = 2c_1.$$

Theorem 9.2 is now applicable.

10. Proof of the theorem for joins

In this section we prove Theorem 4.1 relative to the join models. After some notations, Section 10.1 describes the new bi-dimensional processes related to the total number of balls. Section 10.2 proves a preliminary result (Lemma 2) generalizing Lemma 1 of Section 6. Section 10.3 analyzes the non-random static structure related to the join model. Section 10.4 presents the main theorem, corresponding to the complete random structure.

The function $\varphi(\kappa) = I(\kappa > 0)$ of Section 6 will be denoted hereafter by φ_1. We shall need another function $\varphi(\kappa) = \kappa$, denoted by φ_2. Note that the equijoin (EJ) corresponds to φ_2^R and φ_2^B, and the semijoin (SJ) to φ_2^R and φ_1^B (the functions φ^R are related to the relation R, i.e. to red balls, and similarly for φ^B).

10.1. Processes related to the total number of balls

We need here bi-dimensional processes related to red (R) balls and blue (B) balls. Let us mention for instance:

- *P*8: Here we have red (R) balls and blue (B) balls, furnished by *independent processes*, with probabilities $p_{\mathscr{I}}^R$, $p_{\mathscr{D}}^R$, $p_{\mathscr{Q}}^R$ and $p_{\mathscr{I}}^R$, $p_{\mathscr{D}}^R$, $p_{\mathscr{Q}}^R$ (see Fig. 2). The means and variances corresponding to one step are given by $\bar{x}_R$, σ_R^2, $\bar{x}_B$, σ_B^2 as in $P4$ and the covariance

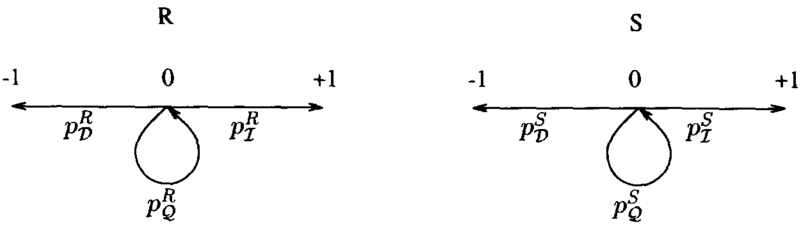

Fig. 2. *P*8 probabilities.

matrix of the bidimensional BM is written as

$$\begin{array}{c} \\ R_s \\ \\ R_t \\ B_s \\ \\ B_t \end{array} \begin{array}{c} \begin{array}{cccc} R_s & R_t & B_s & B_t \end{array} \\ \left(\begin{array}{cccc} \sigma_R^2 s & \sigma_R^2 s & & \\ & & \multicolumn{2}{c}{0} \\ \sigma_R^2 s & \sigma_R^2 t & & \\ & & \sigma_B^2 s & \sigma_B^2 s \\ \multicolumn{2}{c}{0} & & \\ & & \sigma_B^2 s & \sigma_B^2 t \end{array} \right) \end{array} \quad s \leqslant t.$$

Remark. We can of course combine any of the processess P_1–P_7 of Section 5 to furnish *independently* red and blue balls.

- $P9$: We have red (R) balls and blue (B) balls, with probabilities $p_{\mathscr{I}}^R$, $p_{\mathscr{D}}^R$, $p_{\mathscr{I}}^B$, $p_{\mathscr{D}}^B$, $p_{\mathscr{Q}}$ (see Fig. 3), Means and variances corresponding to one step are given by $\bar{x}_R$, σ_R^2, $\bar{x}_B$, σ_B^2 as in $\mathscr{P}_4$ and the covariance matrix of the bidimensional BM is written as

$$\begin{array}{c|cccc} & R_s & R_t & B_s & B_t \\ \hline R_s & \sigma_R^2 s & \sigma_R^2 s & -\bar{x}_B\bar{x}_R s & -\bar{x}_B\bar{x}_R s \\ R_t & \sigma_R^2 s & \sigma_R^2 t & -\bar{x}_B\bar{x}_R s & -\bar{x}_B\bar{x}_R t \\ B_s & -\bar{x}_B\bar{x}_R s & -\bar{x}_B\bar{x}_R s & \sigma_B^2 s & \sigma_B^2 s \\ B_t & -\bar{x}_B\bar{x}_R s & -\bar{x}_B\bar{x}_R t & \sigma_B^2 s & \sigma_B^2 t \end{array} \quad s \leqslant t. \qquad (19)$$

So $f_1^R(t) = \bar{x}_R t, f_1^B(t) = \bar{x}_B t$ and $f_2^{\cdot\cdot}(s, t)$ is immediately written down from (19).

- $P10$: We have red (R) balls and blue (B) with probabilities $p_{\mathscr{I}}^R$, $p_{\mathscr{D}}^R$, $p_{\mathscr{I}\mathscr{I}}^{RB}$, $p_{\mathscr{I}\mathscr{D}}^{RB}$, $p_{\mathscr{I}}^B$, $p_{\mathscr{D}}^B$, $p_{\mathscr{D}\mathscr{I}}^{RB}$, $p_{\mathscr{D}\mathscr{D}}^{BB}$. This process is a generalization of both $P8$ and $P9$; see Figs. 2–4. Set

$$\pi_{\mathscr{I}}^R := p_{\mathscr{I}}^R + p_{\mathscr{I}\mathscr{I}}^{RB} + p_{\mathscr{I}\mathscr{D}}^{RB}, \qquad \pi_{\mathscr{D}}^R := p_{\mathscr{D}}^R + p_{\mathscr{D}\mathscr{I}}^{RB} + p_{\mathscr{D}\mathscr{D}}^{RB}$$

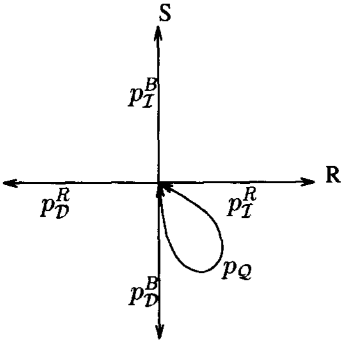

Fig. 3. $P9$ probabilities.

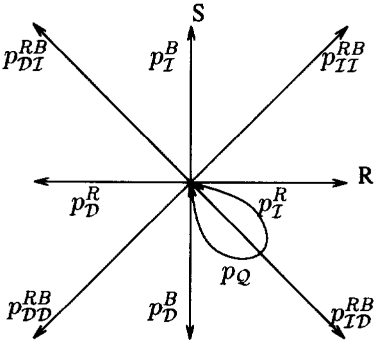

Fig. 4. $P10$ probabilities.

and similarly for π^B_{I}, $\pi^B_{\mathcal{D}}$. Then $\bar{x}_R = \pi^R_{\mathcal{I}} - \pi^R_{\mathcal{D}}$ and the covariance matrix of the bidimensional BM is characterized by $\sigma^2_R = \pi^R_{\mathcal{I}} + \pi^R_{\mathcal{D}} - \bar{x}^2_{\mathcal{R}}$ (similarly for σ^2_B), and

$$f^{R,B}_2(s,t) = COV^{R,B}(s,t) = (p^{RB}_{\mathcal{II}} + p^{RB}_{\mathcal{DD}} - p^{RB}_{\mathcal{ID}} - p^{RB}_{\mathcal{DI}})s - \bar{x}_B\bar{x}_R s, \quad s \leqslant t.$$

- $P11$: We now choose time-dependent probabilities in $P9$. Set $g_{BR}(s) := -\int_0^t \bar{x}_B(s)\bar{x}_R(s)\mathrm{d}s$, and define σ^2 as in the process $P6$ of Section 5. The covariance matrix of the two-dimensional BM is given by

$$\begin{pmatrix} \sigma^2_R(s) & \sigma^2_R(s) & g_{BR}(s) & g_{BR}(s) \\ \sigma^2_R(s) & \sigma^2_R(t) & g_{BR}(s) & g_{BR}(t) \\ g_{BR}(s) & g_{BR}(s) & \sigma^2_B(s) & \sigma^2_B(s) \\ g_{BR}(s) & g_{BR}(t) & \sigma^2_B(s) & \sigma^2_B(t) \end{pmatrix} \qquad s \leqslant t.$$

Time-dependent probabilities can be similarly introduced in $P10$.

10.2. Preliminary result

Let $Y_1 := \sum_{1 \leqslant i \leqslant d} \varphi(\kappa^i_1)\psi(\lambda^i_1)$ for any measurable φ and ψ, where $\kappa^i_1(\lambda^i_1)$ is the number of red (blue) balls at time t_1 in urn U_i. The properties of Y_1 are given by the following lemma (Y_2 is relative to the time t_2):

Lemma 4. *The mean and covariance of Y are given by*

$$E(Y_1) = dE^i_1(\varphi)E^i_1(\psi),$$

$$\begin{aligned} COV(Y_1, Y_2) = {} & dE^{i,i}_{1,2}[\varphi]E^{i,i}_{1,2}[\psi] - dE^{i,j}_{1,2}[\varphi]E^{i,j}_{1,2}[\psi] \\ & + d^2[E^i_1[\varphi]E^j_2[\varphi]C^{i,j}_{1,2}[\psi] + E^i_1[\psi]E^j_2[\psi]C^{i,j}_{1,2}[\varphi] \\ & + C^{i,j}_{1,2}[\varphi]C^{i,j}_{1,2}[\psi]], \end{aligned}$$

where $E_1^i[\varphi]$, $E_{1,2}^{i,j}[\varphi]$ are defined as in Section 6 and with

$$\begin{aligned} C_{1,2}^{i,j}[\varphi] &:= \sum_{k_1} Pr(\kappa_1^i = k_1)\,\varphi(k_1) \sum_{k_2} [Pr(\kappa_2^j = k_2 \mid \kappa_1^i = k_1) - Pr(\kappa_2^j = k_2)]\,\varphi(k_2) \\ &= E_{1,2}^{i,j}[\varphi] - E_1^j[\varphi]E_2^i[\varphi]. \end{aligned}$$

Proof.

$$\begin{aligned} COV &= dE_{1,2}^{i,i}[\varphi]E_{1,2}^{i,i}[\psi] + d(d-1)E_{1,2}^{i,j}[\varphi]E_{1,2}^{i,j}[\psi] \\ &\quad - d^2E_1^i[\varphi]E_2^j[\varphi]E_1^i[\psi]E_2^j[\psi] \\ &= dE_{1,2}^{i,i}[\varphi]E_{1,2}^{i,i}[\psi] - dE_{1,2}^{i,j}[\varphi]E_{1,2}^{i,j}[\psi] - d^2E_1^i[\varphi]E_2^j[\varphi]E_1^i[\psi]E_2^j[\psi] \\ &\quad + d^2(E_1^i[\varphi]E_2^j[\varphi] + C_{1,2}^{i,j}[\varphi])(E_1^i[\psi]E_2^j[\psi] + C_{1,2}^{i,j}[\psi]), \end{aligned}$$

which proves the lemma.

10.3. Nonrandom static structure (NR)

The quantities $E_1^i, E_{1,2}^{i,j}$ for φ_1 are given in Lemma 1 of Section 6.1. We need similar expressions for $\varphi_2(\kappa) = \kappa$. We can derive the following result, which can be found in a more detailed form in [15]:

Proposition 2. *For each join model, there exist two functions $F(\cdot,\cdot)$ and $\psi_{NR}(t_1, t_2)$ such that*

$$E(Y_1) \sim nF(f_1^R(t_1), f_1^B(t_1)),$$

$$COV(Y_1, Y_2) \sim n\psi_{NR}(t_1, t_2).$$

Let us give two examples.

$P9/[\mathscr{A}^R, \mathscr{A}^B]/EJ$: Equijoin, with unbounded urns and probabilities $p_{\mathscr{I}}^R, p_{\mathscr{D}}^R, p_{\mathscr{I}}^B, p_{\mathscr{D}}^B, p_{\mathscr{Q}}$. This leads to

$$E(Y_1) = \alpha n \frac{f_1^R(t_1)}{\alpha}\frac{f_1^B(t_1)}{\alpha} = nF\,(f_1^R(t_1), f_1^B(t_1)) = n\,\frac{\bar{x}_R\bar{x}_Bt_1^2}{\alpha},$$

$$\begin{aligned} &COV(Y_1, Y_2) \\ &\quad \sim dE_{1,2}^{i,i,R}(\mathscr{A},\varphi_2)\,E_{1,2}^{i,i,B}(\mathscr{A},\varphi_2) - dE_{1,2}^{i,i,R}(\mathscr{A},\varphi_2)\,E_{1,2}^{i,j,B}(\mathscr{A},\varphi_2) \\ &\qquad + d^2[E_1^{i,R}(\mathscr{A},\varphi_2)\,E_2^{j,R}(\mathscr{A},\varphi_2)\,C_{1,2}^{i,j,B}(\mathscr{A},\varphi_2) \\ &\qquad + E_1^{i,B}(\mathscr{A},\varphi_2)\,E_2^{j,B}(\mathscr{A},\varphi_2)C_{1,2}^{i,j,R}(\mathscr{A},\varphi_2) + C_{1,2}^{i,j,R}(\mathscr{A},\varphi_2)\,C_{1,2}^{i,j,B}(\mathscr{A},\varphi_2)] \end{aligned}$$

$$\sim n\left\{\frac{1}{\alpha}\left[ps_{1,2}\,f_1(t_1)+\frac{f_1(t_1)f_1(t_2)}{\alpha}\right]^R\left[ps_{1,2}\,f_1(t_1)+\frac{f_1(t_1)f_1(t_2)}{\alpha}\right]^B\right.$$

$$-\frac{1}{\alpha^3}[f_1(t_1)f_1(t_2)]^R[f_1(t_1)f_1(t_2)]^B-\left[f_1^R(t_1)f_1^R(t_2)\,ps_{1,2}^B\frac{f_1^B(t_1)}{\alpha^2}\right]$$

$$\left.-\left[f_1^B(t_1)f_1^B(t_2)ps_{1,2}^R\frac{f_1^R(t_2)}{\alpha^2}\right]\right\}$$

$$\sim\frac{n}{\alpha}ps_{1,2}^R f_1^R(t_1)\,ps_{1,2}^B f_1^B(t_1)=n\bar{x}_R\bar{x}_B t_1^2\,ps_{1,2}^R\,ps_{1,2}^B/\alpha.$$

We use here the notation $[expr]^R$ to indicate that the quantities inside the brackets are the one relative to red balls, and similarly for $[expr]^B$ and blue balls.

- *P9/$[\mathscr{A}^R, \mathscr{B}^B]$/SJ: Semijoin, with unbounded urn for red balls, bounded urns for blue balls and probabilities $p_{\mathscr{I}}^R$, $p_{\mathscr{D}}^R$, $p_{\mathscr{I}}^B$, $p_{\mathscr{D}}^B$, $p_{\mathscr{Q}}$.* This leads to

$$E(Y_1)=\alpha n E_1^{i,R}(\mathscr{A},\varphi_2)E_1^{i,B}(\mathscr{B},\varphi_1)=\alpha n\frac{f_1^R(t_1)}{\alpha}\left[1-\left(1-\frac{f_1^B(t_1)}{\beta}\right)^\delta\right]$$

$$=nF[f_1^R(t_1),f_1^R(t_1)]\sim n\bar{x}_R t_1\left[1-\left(1-\frac{\bar{x}_B t_1}{\beta}\right)^\delta\right],$$

$$COV(Y_1, Y_2)$$

$$\sim dE_{1,2}^{i,i,R}(\mathscr{A},\varphi_2)\,E_{1,2}^{i,i,B}(\mathscr{B},\varphi_1)-dE_{1,2}^{i,j,R}(\mathscr{A},\varphi_2)\,E_{1,2}^{i,j,B}(\mathscr{B},\varphi_1)$$

$$+d^2\,[E_1^{i,R}(\mathscr{A},\varphi_2)\,E_2^{j,R}(\mathscr{A},\varphi_2)\,C_{1,2}^{i,j,B}(\mathscr{B},\varphi_1)$$

$$+E_1^{i,B}(\mathscr{B},\varphi_1)\,E_2^{j,B}(\mathscr{B},\varphi_1)\,C_{1,2}^{i,j,R}(\mathscr{A},\varphi_2)+C_{1,2}^{i,j,R}(\mathscr{A},\varphi_2)\,C_{1,2}^{i,j,B}(\mathscr{B},\varphi_1)]$$

$$\sim n\left\{\left[ps_{1,2}f_1(t_1)+\frac{f_1(t_1)f_1(t_2)}{\alpha}\right]^R\right.$$

$$\times[1-Z_1^i(\mathscr{B},\varphi_1)-Z_2^i(\mathscr{B},\varphi_1)+Z_{1,2}^{i,i}(\mathscr{B},\varphi_1)]^B$$

$$-\frac{f_1^R(t_1)f_1^R(t_2)}{\alpha}[1-Z_1^i(\mathscr{B},\varphi_1)-Z_2^j(\mathscr{B},\varphi_1)+Z_{1,2}^{i,j}(\mathscr{B},\varphi_1)]^B$$

$$-f_1^R(t_1)f_1^R(t_2)\left[\frac{f_1(t_1)}{\alpha^2}f_7(t_1,t_2)\left(1-\frac{f_1(t_1)}{\beta}\right)^\delta\left(1-\frac{f_1(t_2)}{\beta}\right)^{\delta-1}\right]^B$$

$$\left.-\left[1-\left[1-\frac{f_1^B(t_1)}{\beta}\right]^\delta\right]\left[1-\left[1-\frac{f_1^B(t_2)}{\beta}\right]^\delta\right]ps_{1,2}^R f_1^R(t_1)\right\}.$$

10.4. Random structure (*R*)

Now we must analyze the two-dimensional join model. We have two underlying processes with

$$n_1^R = n\left(f_1^R(t_1) + \frac{\theta_1^R}{\sqrt{n}} \right) + \mathrm{O}(1), \qquad n_1^B = n\left(f_1^B(t_1) + \frac{\theta_1^B}{\sqrt{n}} \right) + \mathrm{O}(1)$$

balls at t_1 and

$$n_2^R = n\left(f_1^R(t_2) + \frac{\theta_2^R}{\sqrt{n}} \right) + \mathrm{O}(1), \qquad n_2^B = n\left(f_1^B(t_2) + \frac{\theta_2^B}{\sqrt{n}} \right) + \mathrm{O}(1)$$

balls at t_2, where θ_1^R, θ_2^R, θ_1^B, θ_2^B are Gaussian random variables with mean 0 and covariance $f_2^{\cdot,\cdot}(s, t)$ written down from $P8$–$P11$.

If we fix n_1, n_2, the covariance $COV(Y_1, Y_2)$ is given by $n\Psi_C(t_1, t_2)$ with

$$\Psi_C(t_1, t_2) = \Psi_{NR}(t_1, t_2) + \bar{\varphi}_1^R(t_1, t_2)\frac{\theta_1^R}{\sqrt{n}} + \bar{\varphi}_2^R(t_1, t_2)\frac{\theta_2^R}{\sqrt{n}}$$
$$+ \bar{\varphi}_1^B(t_1, t_2)\frac{\theta_1^B}{\sqrt{n}} + \bar{\varphi}_2^B(t_1, t_2)\frac{\theta_2^B}{\sqrt{n}} + \mathrm{O}\left(\frac{1}{n}\right)$$

for some $\bar{\varphi}$, and Ψ_{NR} is computed in Section 10.3. Set

$$\gamma^R(t_1) := \left. \frac{\partial F(f_1^R, f_1^B)}{\partial f_1^R} \right|_{t=t_1}$$

and similarly for $\gamma^B(t_1)$. We now obtain the following result:

Theorem 10.1. *In the join model, the size* $S([nt])$ *of the join at the time* nt *is asymptotically given by a* (*not necessarily Markov*) *Gaussian process with*

$$E[S([nt])] \sim nG(t) \quad \textit{with } G(t) := F[f_1^R(t), f_1^B(t)],$$

$$COV(S_1, S_2) \sim n\Psi_R(t_1, t_2),$$

with

$$\Psi_R(t_1,\ t_2) := \Psi_{NR}(t_1, t_2)$$
$$+ \gamma^R(t_1)\gamma^R(t_2)f_2^{R,R}(t_1, t_2) + \gamma^R(t_1)\gamma^B(t_2)f_2^{R,B}(t_1, t_2)$$
$$+ \gamma^B(t_1)\gamma^R(t_2)f_2^{B,R}(t_1, t_2) + \gamma^B(t_1)\gamma^B(t_2)f_2^{B,B}(t_1, t_2),$$

$$VAR[S([nt])] \sim n[\Psi_{N,R}(t, t) + \gamma^R(t)^2 f_2^{R,R}(t, t) + 2\gamma^R(t)\gamma^B(t) f_2^{R,B}(t, t)$$
$$+ \gamma^B(t)^2 f_2^{B,B}(t, t)].$$

The relative error in the density due to the asymptotic approximation is $\mathrm{O}(1/\sqrt{n})$.

Proof. We derive

$$
\begin{aligned}
&Ee^{\mathrm{i}[\xi_1 S_1+\xi_2 S_2]}\\
&\quad\sim \exp\Big\{\mathrm{i}\xi_1 nF(f_1^R(t_1), f_1^B(t_1)) + \mathrm{i}\xi_2 nF(f_1^R(t_2), f_1^B(t_2))\\
&\qquad\qquad -\frac{n}{2}[\Psi_{N,R}(t_1,t_1)\xi_1^2 + 2\Psi_{NR}(t_1,t_2)\xi_1\xi_2 + \Psi_{NR}(t_2,t_2)\xi_2^2]\Big\}\\
&\qquad \times E\Big[\exp\Big\{\mathrm{i}\theta_1^R\sqrt{n}\Big[\xi_1\gamma^R(t_1) + \frac{\mathrm{i}}{2}\,\bar\varphi_1^R(t_1,t_1)\xi_1^2 + \mathrm{i}\bar\varphi_1^R(t_1,t_2)\xi_1\xi_2 + \frac{\mathrm{i}}{2}\,\xi_2^2\bar\varphi_1^R(t_2,t_2)\Big]\\
&\qquad\qquad + \mathrm{i}\theta_2^R\sqrt{n}\Big[\xi_2\gamma^R(t_2) + \mathrm{i}\bar\varphi_1^R(t_1,t_2)\xi_1\xi_2 + \frac{\mathrm{i}}{2}\xi_1^2\bar\varphi_2^R(t_1,t_1)\Big]\\
&\qquad\qquad + \mathrm{i}\theta_1^B\sqrt{n}\Big[\xi_1\gamma^B(t_1) + \frac{\mathrm{i}}{2}\,\bar\varphi_1^B(t_1,t_2)\xi_1^2 + \mathrm{i}\bar\varphi_1^B(t_1,t_2)\xi_1\xi_2\\
&\qquad\qquad\qquad + \frac{\mathrm{i}}{2}\xi_1^2\bar\varphi_2^R(t_1,t_2)\Big]\\
&\qquad\qquad + \mathrm{i}\theta_2^B\sqrt{n}\Big[\xi_2\gamma^B(t_2) + \frac{\mathrm{i}}{2}\,\bar\varphi_2^B(t_2,t_2)\xi_2^2 + \mathrm{i}\bar\varphi_1^B(t_2,t_2)\xi_1\xi_2\\
&\qquad\qquad\qquad + \frac{\mathrm{i}}{2}\xi_1^2\bar\varphi_2^B(t_1,t_1)\Big] + \mathrm{O}(1)\Big\}\Big].
\end{aligned}
$$

The last term leads to

$$
\begin{aligned}
\exp\Big\{ &-\frac{n}{2}[\xi_1^2\gamma^R(t_1)^2 f_2^{R,R}(t_1,t_1) + \xi_2^2\gamma^R(t_2)^2 f_2^{R,R}(t_2,t_2)\\
&+ \xi_1^2\gamma^B(t_1)^2 f_2^{B,B}(t_1,t_1) + \xi_2^2\gamma^B(t_2)^2 f_2^{B,B}(t_2,t_2)\\
&+ 2\xi_1\xi_2\gamma^R(t_1)\gamma^R(t_2) f_2^{R,R}(t_1,t_2) + 2\xi_1^2\gamma^R(t_1)\gamma^B(t_1) f_2^{R,B}(t_1,t_1)\\
&+ 2\xi_1\xi_2\gamma^R(t_1)\gamma^B(t_2) f_2^{R,B}(t_1,t_2) + 2\xi_2^2\gamma^R(t_2)\gamma^B(t_2) f_2^{R,B}(t_2,t_2)\\
&+ 2\xi_1\xi_2\gamma^B(t_1)\gamma^R(t_2) f_2^{B,R}(t_1,t_2) + 2\xi_1\xi_2\gamma^B(t_1)\gamma^B(t_2) f_2^{B,B}(t_1,t_2)\\
&+ \text{cubic terms in } \xi_1,\xi_2 + \mathrm{O}(1/\sqrt{n})]\Big\},
\end{aligned}
$$

which proves the theorem. □

Let us illustrate our technique with one simple example (all other cases can be analyzed similarly).

For $P9/[\mathscr{A}^R, \mathscr{A}^B]/EJ$ (equijoin with unbounded urns for both red and blue balls).

$$\gamma^R(t_1) = \frac{f_1^B(t_1)}{\alpha}, \quad \gamma^B(t_1) = \frac{f_1^R(t_1)}{\alpha}$$

and $f_2^{\cdot\cdot}(s,t)$ is directly written down from (19). Theorem 6.1 is now immediately applicable. This leads to

$$E[S([nt])] \sim n\frac{\bar{x}_R\bar{x}_B}{\alpha}t^2,$$

$$COV(S_1,S_2) \sim n\left[\frac{\bar{x}_R\bar{x}_B t_1^2 ps_{1,2}^R ps_{1,2}^B}{\alpha}+\frac{\bar{x}_B^2 t_1 t_2}{\alpha^2}\sigma_R^2 t_1\right.$$
$$\left.-2\frac{\bar{x}_B\bar{x}_R t_1 t_2}{\alpha^2}\bar{x}_B\bar{x}_R t_1+\frac{\bar{x}_R^2 t_1 t_2}{\alpha^2}\sigma_B^2 t_1\right].$$

We should also mention that, if desired, we can characterize the distribution of the maximum size of a join and obtain a result similar to Theorem 9.1, by applying the method of Daniels presented in Section 9 of this paper.

Acknowledgements

We are grateful to C. Lefevre for several discussion on birth and death processes, and to the referees for pertinent comments.

References

[1] P.B. Billingsley, *Convergence of Probability Measures* (Wiley, New York, 1968).

[2] S. Christodoulakis, Estimating block transfers and join sizes, in: Proc. *ACM SIGMOD*, (May 1983) 40–54.

[3] S. Christodoulakis, Implications of certains assumptions in database performance evaluation, *ACM Trans. Database System* 9 (2): (June 1984) 165–186.

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

[5] E. Csaki, A. Foldes and P. Salminen, On the joint distribution of the maximum and its location for a linear diffusion, *Ann. Inst. H. Poincaré* **23** (2) (1987) 179–194.

[6] H.E. Daniels, The maximum of a gaussian process whose mean path has a maximum, with an application to the strength of bundles of fibres, *Adv. in Appl. Probab* **21** (1989) 315–333.

[7] H.E. Daniels, and T.H.R. Skyrme, The maximum of a random walk whose mean path has a maximum, *Adv. in Appl. Probab.* **17** (1985) 85–99.

[8] J. Durbin, The first-passage density of a continuous gaussian process to a general boundary, *J. Appl. Probab.* **22** (1985) 99–122.

[9] P. Flajolet, J. Françon and J. Vuillemin, Sequence of operations analysis for dynamic data structures, *J. Algorithms* (1980) 111–141.

[10] P. Flajolet, C. Puech and J. Vuillemin, The analysis of simple list structures, *Inform. Sci.* **38** (1986) 121–146.

[11] J. Francon and C.Puech, Histoires de files de priorité avec fusions, in: B. Courcelle, ed., *CAAP'84, 9th Colloq. on Trees in Algebra and Programming, Cambridge University Press*, Bordeaux (France), March 1984).

[12] D. Gardy, Normal limiting distributions for projection and semijoin sizes, *SIAM J. Discrete Math.* **5** (2) (1992) 219–248.

[13] D. Gardy, Join sizes, urn models and normal limiting distributions, *Theret. Comput. Sci.* **131** (1994) 375–414.

[14] D. Gardy and G.Louchard, Dynamic analysis of some relational data base parameters I: projections, Tech. Report 94–6, Lab. Prism, University of Versailles, February 1994.

[15] D. Gardy and G. Louchard, Dynamic analysis of some relational data base parameters II: equijoins and semijoins, Tech. Report 94–7, Lab. Prism, University of Versailles, February 1994.

[16] D. Gardy and C.Puech, On the sizes of projections: a generating function approach, *Inform Systems* **9** (3/4) (1984) 231–235.

[17] D. Gardy and C. Puech, On the effect of join operation on relation sizes, *ACM Trans. Database Systems* **14** (4) (1989) 574–603.

[18] P.J. Haas, J.F. Naughton, S. Seshadri, and A.N. Swami, Fixed-precision estimation of join selectivity, in: *Principles of Database Systems* (Washington, 1993) 190–201.

[19] W.C. Hou and G. Ozsoyoglu, Statistical estimators for aggregate relational algebra queries, *ACM Trans. Database System* **16** (4) (1991) 600–654.

[20] N.L. Johnson and S. Kotz, *Urn Models and Their Application* (Wiley, New York, 1977).

[21] S. Karlin and H.M. Taylor, *A Second Course in Stochastic Processes* (Academic Press, New York, 1981).

[22] C.M. Kenyon-Mathieu and J.S. Vitter, General methods for the analysis of the maximum size of dynamic data structures, in: *Proc. 16th Internat. Colloqu. on Automata, Languages and Programming,* Stresa, Italy, July 1989, Lecture Notes in Computer Science, Vol. 372 (Springer, Berlin) 473–487.

[23] Y. Ling and W. Sun, A supplement to sampling-based methods or query size estimation in a database system, *SIGMOD Record* **21** (4) (1992).

[24] R.L. Lipton, J.F. Naughton, D.A. Schneider and S. Seshadri, Efficient sampling strategies for relational database operations, *Theoret. Comput. Sci.* **116** (1) (1993) 195–226.

[25] G. Louchard, Brownian motion and algorithms complexity, *BIT* **26** (1986) 17–34.

[26] G. Louchard, Random walks, gaussian processes and list structures, *Theoret. Comput. Sci.* **53** (1987) 99–124.

[27] G. Louchard, Trie size in a dynamic list structure, in: M.-C. Gaudel and J.-P. Jouannaud, eds., *Proc. TAPSOFT'93*.

[28] G. Louchard, R. Schott and B. Randrianarimananaa, Dynamic algorithms in D.E. Knuth's model: a probabilistic analysis, *Theoret. Comput. Sci.* **93** (1992) 201–225.

[29] R.S. Maier, A path integral approach to data structure evolution, *J. Complexity* (1991) 232–260.

[30] M.V. Mannino, P. Chu and T. Sager, Statistical profile estimation in database systems, *ACM Comput. Surveys* **20** (3) (1988) 191–221.

[31] T.H. Merrett and E. Otoo, Distribution models of relations, in: *Proc. 5th Conf. on Very Large Data Bases* (Rio de Janeiro, October 1979) 418–425.

[32] J.K. Mullin. Estimating the size of a relational join, *Inform. Systems* **18**(3) (1993) 189–196.

[33] B. Muthuswamy and L. Kerschberg, A detailed statistical model for relational query optimization, in: *Proc. Ann. Conf. of the ACM* (Denver, Colorado, October 1985) 439–448.

[34] G. Piatetsky-Shapiro and C. Connell, Accurate estimation of the number of tuples satisfying a condition, in: *Internat. Conf. ACM SIGMOD* (Boston, June 1984) 256–276.

[35] W. Sun, Y. Ling, N. Rishe and Y. Deng, An instant and accurate size estimation method for joins and selection in a retrieval-intensive environment, in: *Proc. ACM SIGMOD Internat. Conf.* (Washington, DC, May 1993) 79–88.

[36] A. van Gelder, Multiple join size estimation by virtual domains, in: *Principles of Database Systems* (Washington, DC, 1993) 180–189.