capstract:2610.00005v1 [math.PR] 27 Oct 2026

Pedigree collapse in a finite population of presentation files

Léonie S. Verhagenne[1,*] Darío V. Lencastre[2]
[1]Section of Applied Probability, Meerdal Institute of Mathematics, Breda, Netherlands
[2]Department of Mathematical Sciences, Instituto de Estudios de Valdebruna, Zaragoza, Spain

October 27, 2026

Abstract

A revision of a slide presentation can have two recorded parents, the file opened for duplication and the donor from which the editor imports a set of slides. We consider the studbook obtained by entering both parents even when the editor subsequently deletes their contributions. At each of two positions in the offspring record we transmit one of the corresponding parent's two ancestral marks, with equal probabilities. For N files per revision and a probability q of deliberately using the same file for both operations, we obtain an exact law for the first common mark. Its mean depth within a file is $2 + 2(N-1)(1-q)$ revisions. When deliberate reuse occurs with probability $1 - c/N$, the depth converges in distribution to a geometric variable of mean two, although its expectation tends to $2 + 2c$. The discrepancy comes from occasional imports that require a search across the entire archive. A worked register gives an inbreeding coefficient of one eighth for a municipal presentation whose parents share a retired briefing on the projector remote.

Keywords: identity by descent, partial selfing, coalescence time, file provenance, finite Markov chain

1 Introduction

The computer from which a municipal officer presents the annual accounts usually contains an earlier presentation of the annual accounts. It may also contain the induction session, a snow clearance return and the deck that used to explain the projector remote. When the officer opens one file, saves a copy under a new name and imports the slides of a second, the new file has two quite specific relations of descent. The first predecessor supplied the container. The second supplied the imported packet. Subsequent editing can remove every visible trace of either transaction. A register kept at the time of the transactions can nevertheless establish that the two predecessors had the same father.

We study a studbook kept on this principle. An entry identifies the two parent accessions of a presentation, in order. A father in the examples is the file in the first column, irrespective of which parent supplied the substance of the presentation. A full sister has the same two entries in the same order. These conventions come from the register described by Melcor and Thervenne [11], which began with a collection of projection materials catalogued by consecutive accession. The alternative convention of listing the larger contributor first [1] makes the sex of a parent de-

∗Corresponding author: l.verhagenne@meerdal.example

pend on whether the secretary kept the appendix. None of the probability results requires this further information.

Descent of a document has usually been established from features of the document itself. Variants shared between manuscripts are the evidence for much of textual stemmatics [10]. Histories of revision kept at the point of copying give a different kind of evidence [15, 19]. In the archive considered here the path from a dissertation on winter gritting to its grandfather can be recovered without having to find a reference to rock salt in the grandfather. The question is what to count when the grandfather occurs on both sides. A table of distinct ancestors does not record how often the paths cross the same accession or which of its two entries the paths then follow.

We use the identity-by-descent coefficient familiar from pedigree analysis [17, 20]. Imagine two marks in the accession record for each founder. For a new file one mark is sampled from each parent. The marks have no visible expression. Conditional on the studbook, the probability that the two marks in a file trace back to the same mark is its inbreeding coefficient. The probability model on studbooks allows us to ask how this coefficient varies with the size of the file pool and with the rate at which the editor duplicates a presentation and takes the imported slides from that very presentation. Both operations can occur during a single session. The relevant event is their agreement on an accession, including an agreement reached by chance.

The finite archive is a diploid partially selfing population at a single neutral locus. Versions of the ancestral process for such populations occur in [12, 13]. Our use of that calculation is quite literal. It specifies which register entries to request if, for instance, the pensions committee requires a certificate stating the degree of consanguinity of its quarterly briefing. A finite register stops at a definite date. We therefore retain the complete distribution of the depth of the first common mark, including the first few revisions which disappear upon rescaling in a coalescent limit. Of particular interest is a filing practice in which one outside import per generation is about the usual number, even in a large archive. The law of a certificate drawn under this practice is nearly the law under exclusive internal copying. Its mean retrieval depth can be twice as large.

2 The register and its probability law

An archive is divided into revision rounds indexed by integers. Each round contains $N \geq 2$ accessioned files. Taking a copy creates a new accession even if the operator accepts the suggested file name and makes no other change. Paths in the register go from a child at round r to a parent at round $r-1$. Dormant files have to be carried forward explicitly to fit this convention. In a register with actual dates, several edits completed before a committee session could constitute one round. We make no identification between a round and a fixed number of calendar days.

The first parent S is the file duplicated to create the child. The second parent D is the declared donor of the slide packet. For each child, draw S uniformly from the preceding round. With probability q put $D = S$. With the remaining probability draw D uniformly from all N files, including S. Use independent draws for different children and rounds. An editor who takes both sources from a list of recently opened files may return to the first source without intending a same-file import. The probability of a common accession, denoted by h, is consequently

$$h = q + (1 - q)/N. \tag{1}$$

This distinction matters in small offices. With four eligible files even an editor who always consults the list afresh will have a common accession on a quarter of the entries. The excess above

one quarter has a direct interpretation in Eq. (1). Removing the first accession from the list before the second draw would define a different model. Sampling descriptions which omit the treatment of the first accession are particularly difficult to compare [3].

At a founding round put two different marks on the register entry of each file, using $2N$ marks in all. At each later entry give the left position a uniformly sampled mark from S and the right position a uniformly sampled mark from D. Conditional on the pedigree these index choices are independent for all positions. Drawing twice from the same parent can copy the same position twice. It can also copy two different positions whose marks have already come to agree. The mark is a device for averaging over paths of descent and need never be written to disk. It has no interpretation as a master layout identifier, since a layout generally passes entire from one source.

For the two positions of a present file follow their register pointers backward. Let T be the first positive number of rounds after which the pointers occupy the very same position. From that time on they follow one pointer. The finite-depth coefficient F_t is the probability that $T \le t$, with all earlier entries effectively treated as unrelated founders at depth t. An unknown parent should not be entered as a common accession called Unknown. Such a practice introduces a particularly prolific founder and gives its progeny a computed degree of consanguinity which cannot be recovered from the records [16].

We use two levels of conditioning. Given the list of parents, the only random quantities in the marks are choices between two positions. With a random list of parents and a random present file, F_t also averages over source choice. A register with all its parents specified can therefore produce a coefficient larger than F_t, including in a round in which no editor went back to the duplicated file for slides. The conditional calculation is given in Section 5. It is the one appropriate to an individual briefing.

3 Backward transitions

Prior to the first common position there are two possible arrangements of a pair of pointers. Write A when they occupy different files and B when they occupy both positions of the same file. A third state C records absorption. Sampling the two positions of a new deck puts the process in B. Drawing one position from each of two distinct decks puts it in A, even if later inspection should show that the decks were siblings. These starting conditions give different answers at modest depth.

Lemma 1 (Pair kernel)**.** *In the order A, B, the substochastic transition matrix on unmerged pairs is*

$$Q = \begin{pmatrix} 1 - 1/N & 1/(2N) \\ 1 - h & h/2 \end{pmatrix}. \tag{2}$$

The missing probability in each row is the probability of entering C.

Proof. For pointers in different files the two selected parents are independent uniform accessions. This holds even when the pointers are known to occupy the same column of their respective records. If their parents coincide, two independent uniform position choices agree with probability one half. For pointers in the same file the two parents are its two recorded accessions. They coincide with probability h. The position draws again agree with probability one half given that coincidence. Up to the first visit to a round its parent records are independent of the pointers which brought the search there. This gives the two rows at every step. ∎

Let e_A, e_B be row vectors and let $\mathbf{1}$ be a column of ones. The survival probability from state i is $e_i Q^t \mathbf{1}$, including at $t = 0$. Thus a finite-depth register has the mean within-file coefficient

$$F_t = 1 - e_B Q^t \mathbf{1}. \tag{3}$$

The first two coefficients are $h/2$ and $h/2 + (1-h)/(2N) + h^2/4$. The middle term in the latter expression allows two pointers that separated into different files on the first step to coalesce on the second. Leaving it out would make all subsequent unions depend exclusively on the pedigree of the duplicated file. The general use of substochastic kernels to sum entrance times is described in [7, 14].

In Figure 1 the archive contains sixteen files in each round. With a half chance of directing the import back to the duplicate, just over half the presentations have identical marks by depth eight, in the joint law of the register and marks. In an actual archive each presentation has a conditional probability of identity. At depth one there are only two possible such probabilities, zero and one half. A histogram of individual coefficients would have neither the ordinate nor the sampling interpretation of the figure.

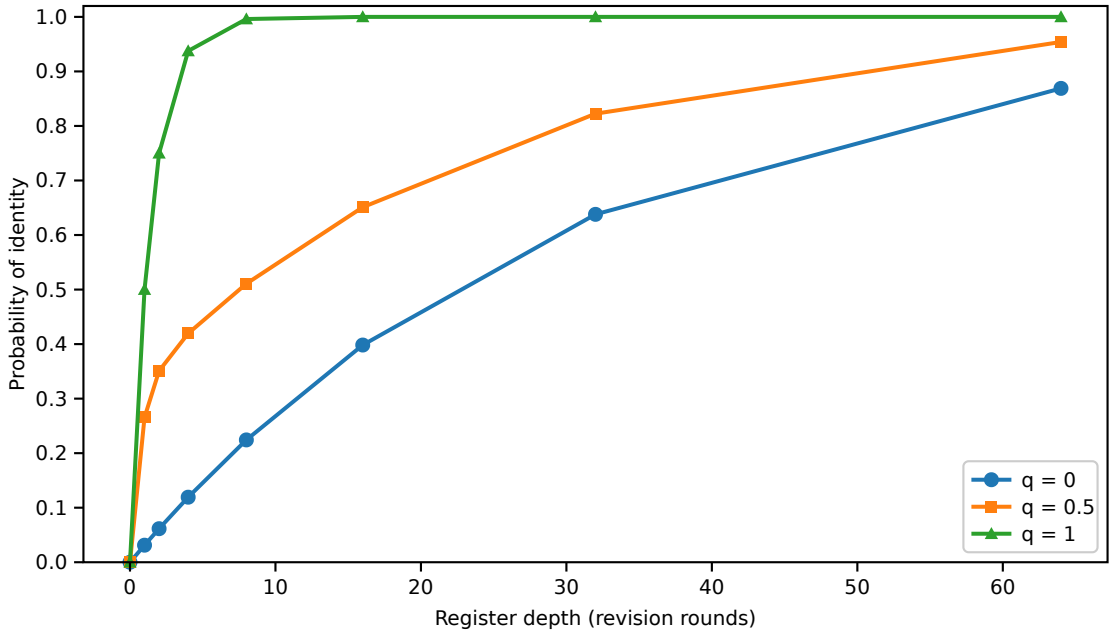


Figure 1: Mean within-file identity by register depth for sixteen files per round. Each point is an evaluation of the pair recursion.

Theorem 2 (Mean depth of a certificate)**.** *Let m_i be the mean number of rounds to the first common position in an archive extended indefinitely into the past, beginning in state i. Then*

$$m_B = 2 + 2(N-1)(1-q), \qquad m_A = N + m_B/2. \tag{4}$$

Proof. The conditional absorption probability on each step of Lemma 1 is at least $1/(2N)$. Both expectations are finite. First-step analysis gives $m_A = 1 + (1 - 1/N) m_A + m_B/(2N)$ and $m_B = 1 + (1-h) m_A + h m_B/2$. The first equation yields the stated expression for m_A. Substituting it into the second yields $m_B = 2 + 2N(1-h)$. Use Eq. (1). ∎

The lower bound on the one-step chance of absorption gives a convenient geometric upper bound for the tail. It is attained throughout when $q = 0$. In that case T has mean $2N$ in either starting state. At $q = 1$, the tail starting from B is exactly 2^{-t} for every archive size. Once both pointers occupy a single file, subsequent searching requires at most one file per round. If the search begins with two different same-source presentations, the archivist must first locate a shared file. Its mean waiting time is N. With probability one half an additional geometric search begins when the file is located, which gives the $N + 1$ in Eq. (4) at this endpoint.

4 Infrequent outside imports

Consider a sequence of archives with $q = 1 - c/N$ for fixed $c > 0$ and $N \geq c$. An expected c children in each round receive an independent donor draw. A few of these draws return to the first parent. Denote the depth of the certificate started within a file by T_N. A common approximation would replace almost exclusive use of the primary parent by exclusive use throughout the search. This replacement is accurate for the percentage of certificates issued within any fixed number of rounds. Its error in the mean persists.

Proposition 3 (A finite mean contribution from escaped pairs)**.** *As N increases, T_N converges in distribution to a geometric variable on the positive integers with parameter $1/2$. The limit of its means is $2 + 2c$. The probability of visiting different files before coalescence is $(1 - h)/(1 - h/2)$.*

Proof. Until the process leaves state B it repeats there with probability $h/2$ at each step. The number of steps L to its first exit is geometric with parameter $1 - h/2$. The exit goes to A with probability $(1 - h)/(1 - h/2)$ and otherwise goes directly to C. This exit type is independent of L. Since $1 - h = c(1 - 1/N)/N$, the chance of reaching A tends to zero, and L converges in distribution to the stated law. On paths avoiding A, $T_N = L$, which proves distributional convergence. The expectation limit follows from Theorem 2. ∎

Each escape requires an additional mean m_A steps by the Markov property. In this regime m_A/N tends to one, and N times the escape probability tends to $2c$. Their product accounts for the missing $2c$ rounds. The family of depths is not uniformly integrable, in the sense of [2]. This example does not require a heavy-tailed input distribution. All random choices have finite support and for every fixed archive there is a geometric bound on certificate length. It is the dependence of that bound on archive size that permits the discrepancy.

Table 1 gives finite values with one independent draw per round in expectation. For a collection of 256 presentations, the entry at eight rounds is a survival probability of about one in 88. The idealized same-file law would give one in 256. Both would place the ordinary certificate near the front of the accession book. One occasionally needs the previous volume.

Table 1: Tail probabilities and mean depth for one independent donor draw per round in expectation. Probabilities come from the two-state matrix, and depths are in rounds.

Files per round	Unresolved after 2	Unresolved after 8	Unresolved after N	Mean depth
8	0.34955	0.10785	0.10785	3.75000
16	0.30590	0.08317	0.05010	3.87500
32	0.27957	0.05311	0.02405	3.93750
64	0.26520	0.03135	0.01176	3.96875
128	0.25771	0.01840	0.00582	3.98438
256	0.25388	0.01136	0.00289	3.99219

The slow scale can also be identified without the general machinery for coalescent convergence [8, 13]. From A, the number of steps to the next encounter in a single file has a geometric distribution with parameter $1/N$. On encounter the two positions agree with probability one half. Otherwise the process enters B, where it makes a geometrically bounded number of additional steps before it either coalesces or escapes again. The probability of escape from this visit is $O(1/N)$ by Proposition 3. Thus, started from A, the total remaining time divided by N converges in law to a unit exponential. Indeed the first waiting time has this limit, the additional visit to B divided by N vanishes in probability, and the event of a second escape has vanishing probability. This argument is a direct coupling to the first encounter clock [9].

For any $s > 0$ these observations give a more precise tail for the search begun in B:

$$N \Pr(T_N > \lfloor N s \rfloor) \longrightarrow 2c \exp(-s). \tag{5}$$

To see this, split at the first exit as in the proof of Proposition 3. The contribution from $L > \lfloor N s \rfloor$ is bounded by $2^{-\lfloor N s \rfloor}$ before multiplication by N. Conditional on escape, L/N tends to zero and the remaining time divided by N has the continuous unit exponential limit just obtained. Multiplication by N times the escape probability gives Eq. (5). The restriction to positive s excludes the initial layer during which almost every certificate is issued. Figure 2 uses $s = 1$ and the unrounded entries underlying Table 1.

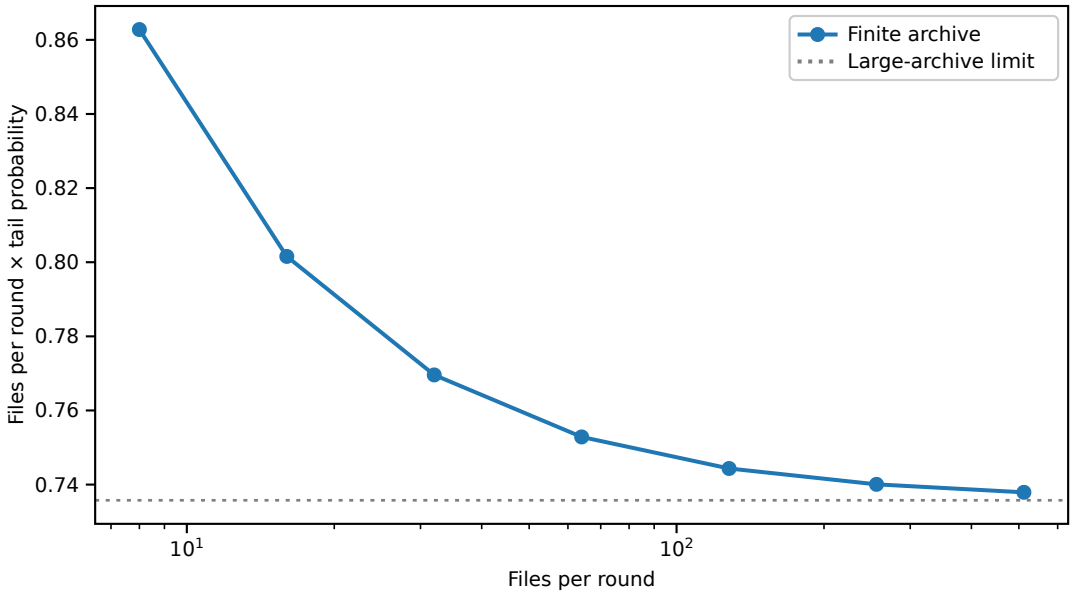


Figure 2: Scaled probability that a certificate requires more rounds than there are files per round, with one independent donor draw per round in expectation. The dotted line is twice the reciprocal of Euler's number.

5 Coefficients for a specified studbook

For an archivist holding a parent list, the probability q of intentional same-source use has no part in the calculation. Define $K(x, y)$ to be the probability that positions sampled independently with replacement from x and y carry the same mark, conditional on the list. For two distinct founders $K(x, y) = 0$, and for a founder x we have $K(x, x) = 1/2$. The inbreeding coefficient $f(x)$ compares the two distinct positions inside x. It is zero for founders. The difference between drawing with replacement and drawing distinct positions is relevant even in the first volume of the register.

For a nonfounder x write s_x, d_x for its two parents. A parent contributes one random position. Accordingly

$$f(x) = K(s_x, d_x), \qquad K(x, x) = (1 + f(x))/2. \tag{6}$$

For a distinct accession y in the same or an earlier round than x,

$$K(x, y) = \tfrac{1}{2}\{K(s_x, y) + K(d_x, y)\}. \tag{7}$$

For Eq. (7), condition on which of the two positions in x was sampled. Its pointer choice is independent of the position drawn from y and of all the earlier pointer choices that determine the marks on the parent. This remains true if x and y were made from exactly the same two files. The case $x = y$ is excluded because the two draws from x can select a single already transmitted position; this is precisely the first half of the diagonal term in Eq. (6). Symmetry of K allows these identities to be used in either order in a round. They are the ordinary tabular pedigree calculation [5]. No independent mark draws are required in its implementation.

Consider the small register in Table 2. There are three founders, all entered on the occasion of a change of projection equipment. The projector remote briefing is the first recorded parent of both the snow clearance deck and the pension briefing. All diagram labels and diagrams from the remote briefing may by this time have been deleted. The question on the certificate is whether snow clearance and pensions are close enough relatives to give their joint report a positive inbreeding coefficient. The clerk can answer it on the basis of the five earlier accession numbers.

Table 2: A register excerpt. A blank in a founder’s parent column means that tracing ends at that accession. Sibling relationships use the ordered parent columns.

Accession	Subject on cover	Duplicated file	Imported packet from	Within-file coefficient
01	Projector remote			0
02	Fleet depot			0
03	Actuarial return			0
11	Snow clearance	01	02	0
12	Municipal pensions	01	03	0
21	Joint quarterly report	11	12	1/8
22	Winter salt allocation	11	11	1/2

To compute the coefficient for accession 21, one pointer must come from the 01 side of accession 11 and the other from the 01 side of accession 12. Each event has probability one half. Conditional on both events, the two positions independently transmitted earlier from the projector briefing match with probability one half. The product is one eighth. For 22 the immediate parents coincide and have two different marks, hence the probability is one half. The use of the same snow clearance deck twice has been recorded as such even though the imported packet might have consisted of only its final slide.

The conditional calculation gives more information than the degree of consanguinity on a single certificate. By Eq. (7) the relationship between two distinct reports from the same pair of unrelated, noninbred files has kinship one quarter. If the two reports have one unrelated, noninbred parent in common and all other founders are distinct, it is one eighth. The diagonal entry for any one of these reports is one half. Figure 3 plots the kinship matrix for the four nonfounders in Table 2 together with the remote briefing. In particular the winter salt allocation has diagonal entry three quarters even if the salt slides themselves contain no repeated words.

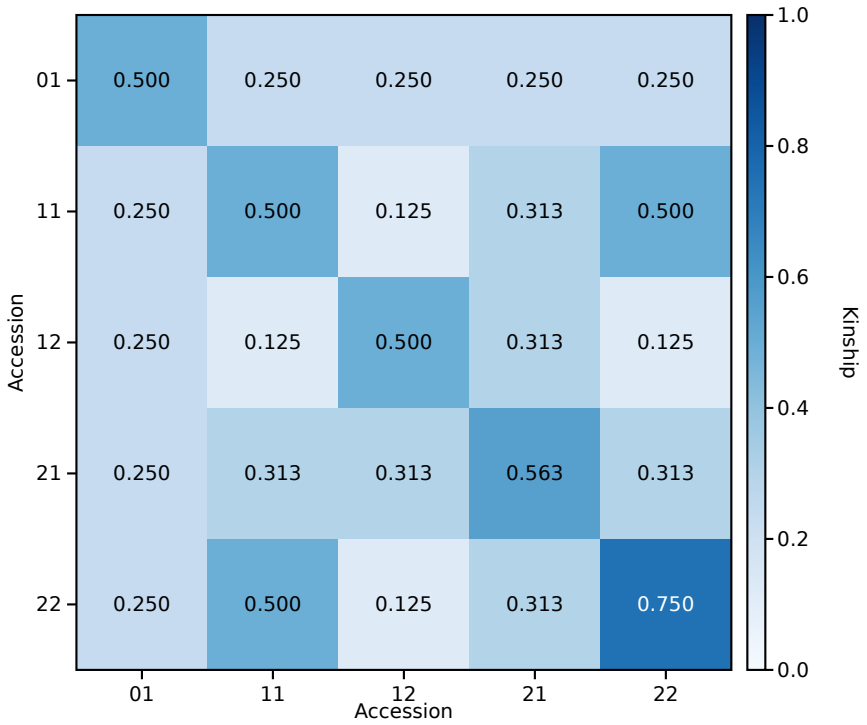


Figure 3: Conditional kinship between accessions in the worked register. Rows and columns denote files, the diagonal samples with replacement, and the numbers are probabilities rounded to three decimals.

There is a straightforward audit for the retrospective assignment of parents. If the clerk cannot determine whether the remote briefing or the fleet depot deck was duplicated, and knows only that the other one supplied the import, the register has two possible orientations for this union. Interchanging the columns of a given union does not affect its own coefficient, by symmetry in Eq. (6). It does not affect the coefficient of any descendant either. For the latter claim, induction in accession order applies Eq. (7) to off-diagonal terms and then Eq. (6) to diagonals. Thus missing orientation can prevent the clerk from certifying paternal half siblings without preventing certification of the degree of inbreeding of their issue.

6 What can be learned from the projection

Whether a file contains the drawing of the remote control pointing at the ceiling is a different observation from whether a pointer reaches accession 01. Suppose that, after the choice of the two parents and before deposit of the file, each slide has been replaced by a new slide supplied by the current speaker. Keep the parent transactions as accessioned. Replacing all displayed data by these fresh slides is compatible with every mark assignment on the given studbook. If the fresh slides are drawn independently of file ancestry from the same law at each accession, all possible studbooks have the same likelihood from this final projection alone. Thus an exact certificate and a wholly uninformative visual inspection can be issued at the same time.

This loss of information can be stated without assumptions on the contents of real slides. Construct two fixed registers for the same seven titles. The first is Table 2. In the second, replace the primary parent of accession 12 by 02 and its secondary parent by 03. The two intermediate

reports still share one parent, so this particular change leaves the coefficient of 21 at one eighth and moves its shared grandfather to the fleet depot. To change the coefficient itself, give 12 two copies of 03 as parents. Then 11 and 12 have disjoint founder sets, and the coefficient of 21 is zero. Under complete replacement by independent current material, each register can generate the identical distribution of files at projection. Even a perfect transcript of the speaker cannot discriminate among them in that experiment.

In practice an investigator might have the package history, the imported master or a memorandum identifying one donor [4]. These observations can restrict the set of feasible registers. File size is a less specific witness. A large inherited photograph may occupy most of the package even after the editor has removed it from every projected slide [18]. The registrar who assigns parentage in proportion to the number of bytes copied must specify which size is measured. The symmetric mark convention avoids this choice at the cost of answering a narrower question. It counts routes in the register. A report with a coefficient of one eighth has no prescribed fraction of old material.

The neutral draw over parents in Section 2 is also a substantive restriction. A directorate may circulate exactly one approved master every year, causing selection probabilities to differ sharply between the two parental roles. The transition out of A would then depend on the columns occupied by the two pointers, and it would generally retain information about the identity of a parent beyond the first step. Sampling complete pedigrees from an existing archive poses a separate problem: files called Final are more likely to survive the annual deletion than files called Copy of Final [6]. The calculation for a surviving, documented family remains meaningful. Inferring N or q from these survivors using Eq. (3) requires a sampling model.

One use of the finite-depth answer is to specify precisely what a certificate can establish when the catalogue opens halfway through the family. If two paths are already united in the available register, earlier accessions cannot separate them. If they end in distinct positions at the boundary, an earlier union is still possible. Treating all boundary positions as founders therefore gives a lower bound for identity by descent over a longer history, path by path. The date of the opening volume belongs on a certificate. Two municipal departments can issue different coefficients for the same presentation simply because one has kept the records from the year the remote was installed.

A Arithmetic and checks on the register

For the finite calculations let a_t and b_t denote the unconditional probability of occupying transient states A and B at depth t. Start with $(a_0, b_0) = (0, 1)$ for the plots of a within-file certificate. The row recursion from Eq. (2) gives $a_{t+1} = (1 - 1/N)a_t + (1 - h)b_t$ and $b_{t+1} = a_t/(2N) + hb_t/2$. The surviving mass decreases on each step by $a_t/(2N) + hb_t/2$, which also happens to equal b_{t+1}. This second identity supplies a check independent of the matrix powers: the mass entering a common position must equal the mass newly assigned to a common file with distinct positions, since an encounter chooses between these two outcomes with a fair draw.

For $N = 16$ and $q = 1/2$, the first step gives $a_1 = 15/32$ and $b_1 = 17/64$. The survival probability is $47/64$. After the second step the probability absorbed is $17/64 + 15/1024 + 289/4096 = 1437/4096$. This is the second nonzero point of the middle curve in Figure 1 after rounding. Probabilities in the table of tail depths were obtained by summing the transient row, including at depths at which its mass is small. Subtracting a rounded value of F_t from one can hide the last surviving families.

For the kinship plot use rows 01, 11, 12, 21 and 22 of Table 2. As an off-diagonal check, $K(21, 22)$ is the average of $K(11, 22)$ and $K(12, 22)$. The former is one half, since 22 has 11 for both parents. The latter equals the kinship of the two half siblings, one eighth. The plotted entry is therefore five sixteenths. This quantity and the diagonal value $K(21, 21) = 9/16$ would both be larger if the clerk had entered the same founder under 02 and 03. A difference between two names for the depot would then survive in the certificate after the last reference to either name had been deleted from the presentation.

Acknowledgments

We thank Alina Verroux for pointing out that the order of the parent columns could be lost while the coefficient remained recoverable.

References

[1] Lucile B. Agerin. Ordering Contributors in Registers of Municipal Lecture Material. *Transactions on Local Record Practice*, 22:103–119, 2019.

[2] Patrick Billingsley. *Convergence of Probability Measures*. Wiley, 1999.

[3] Marín Borrel-Medern. Sampling Donor Histories from Local Projection Collections. *Bulletin of Documentary Inference*, 4:22–39, 2021.

[4] Elise M. Dalcourt-Viens. The Evidence of Covering Memoranda in Hybrid Presentation Records. *Quarterly Review of Archival Methods*, 24:62–79, 2020.

[5] L. O. Emik and C. E. Terrill. Systematic Procedures for Calculating Inbreeding Coefficients. *Journal of Heredity*, 40:51–55, 1949.

[6] Maëlle Isbert-Laroche and Tuva R. Senvik. Survival of Named Revisions at Annual Disposal. *Transactions on Local Record Practice*, 27:111–130, 2024.

[7] John G. Kemeny and J. Laurie Snell. *Finite Markov Chains*. D. Van Nostrand, 1960.

[8] J. F. C. Kingman. The Coalescent. *Stochastic Processes and their Applications*, 13:235–248, 1982.

[9] Torgny Lindvall. *Lectures on the Coupling Method*. Wiley, 1992.

[10] Paul Maas. *Textual Criticism*. Clarendon Press, 1958.

[11] Eva L. Melcor and Noé Thervenne. An Accession Register for Composite Materials in Projection Archives. *Studies in Administrative Collections*, 9:81–108, 2016.

[12] Magnus Nordborg. Linkage Disequilibrium, Gene Trees and Selfing: An Ancestral Recombination Graph With Partial Self-Fertilization. *Genetics*, 154:923–929, 2000.

[13] Magnus Nordborg and Peter Donnelly. The Coalescent Process With Selfing. *Genetics*, 146:1185–1195, 1997.

[14] J. R. Norris. *Markov Chains*. Cambridge University Press, 1997.

[15] Enio Ravelot and Alma Zeffren. A Register of Revision Events for Transient Documents. *Quarterly Review of Archival Methods*, 18:56–73, 2014.

[16] Anja G. Tarsen and Luís Dellorim. Unresolved Accessions in the Enumeration of Descent Paths. *Studies in Administrative Collections*, 13:201–219, 2020.

[17] Elizabeth A. Thompson. Identity by Descent: Variation in Meiosis, Across Genomes, and in Populations. *Genetics*, 194:301–326, 2013.

[18] Henrik Torven-Kall. Unreferenced Objects in Deposited Slide Packages. *Bulletin of Documentary Inference*, 7:33–49, 2024.

[19] Irene M. Velser-Garau. Transaction Witnesses in the Custody of Office Files. *Records of Contemporary Administration*, 6:14–32, 2011.

[20] Sewall Wright. Coefficients of Inbreeding and Relationship. *The American Naturalist*, 56:330–338, 1922.