Probability in the Engineering and Informational Sciences, **28**, 2014, 83–99.
doi:10.1017/S026996481300034X

GENERALIZED BIRTHDAY PROBLEMS IN THE LARGE-DEVIATIONS REGIME

M. MANDJES
Center for Mathematics and Computer Science
Amsterdam, The Netherlands
E-mail: M.R.H.Mandjes@uva.nl

EURANDOM
Center for Mathematics and Computer Science,
Amsterdam, The Netherlands;
EURANDOM, Eindhoven University of Technology,
Eindhoven, The Netherlands

This paper considers generalized birthday problems, in which there are d classes of possible outcomes. A fraction f_i of the N possible outcomes has probability α_i/N, where $\sum_{i=1}^d f_i = \sum_{i=1}^d f_i\alpha_i = 1$. Sampling k times (with replacements), the objective is to determine (or approximate) the probability that all outcomes are different, the so-called *uniqueness probability* (or: *no-coincidence probability*). Although it is trivial to explicitly characterize this probability for the case $d = 1$, the situation with multiple classes is substantially harder to analyze.

Parameterizing $k \equiv aN$, it turns out that the uniqueness probability decays essentially exponentially in N, where the associated decay rate ζ follows from a variational problem. Only for small d this can be solved in closed form. Assuming α_i is of the form $1 + \varphi_i\varepsilon$, the decay rate ζ can be written as a power series in ε; we demonstrate how to compute the corresponding coefficients explicitly. Also, a logarithmically efficient simulation procedure is proposed. The paper concludes with a series of numerical experiments, showing that (i) the proposed simulation approach is fast and accurate, (ii) assuming all outcomes equally likely would lead to estimates for the uniqueness probability that can be orders of magnitude off, and (iii) the power-series based approximations work remarkably well.

1. INTRODUCTION

The classical birthday problem [5,16] can be formulated as follows. Consider an experiment with $N \in \mathbb{N}$ possible outcomes, all with probability $1/N$. Now sample $k \in \{1, \ldots, N\}$ times (with replacements), where subsequent samples are independent. The question is: what is the probability $\pi(k, N)$ that all outcomes are different? This *uniqueness probability* (or: *no-coincidence probability*) can be computed in an elementary way, and it is well-known that the outcome is

$$\pi(k, N) = \frac{N}{N}\frac{N-1}{N}\cdots\frac{N-k+1}{N} = \frac{N!}{(N-k)!N^k}. \tag{1}$$

© Cambridge University Press 2013 0269-9648/14 $25.00

https://doi.org/10.1017/S026996481300034X Published online by Cambridge University Press

The birthday problem turns out to be useful in many application domains, ranging from biology [6] to coding [19]. Interestingly, for some applications the uniqueness probability should be high (think of coding), whereas in others it should be kept low. An example of the latter type of applications could relate to the domain of privacy and security—when one fills out a questionnaire, and reveals age and zip code, one would like to be sure that the uniqueness probability (to be interpreted as the probability that he or she can be uniquely traced) is sufficiently low.

A complication lies in the fact that in many real-life situations the probability distribution is *not* uniform; instead there are probabilities p_1 up to p_N (summing to 1). It is not hard to realize that the uniqueness probability can now be expressed in terms of elementary symmetric functions:

$$\pi(k,N) = k! \cdot \left(\sum_{i_1<i_2<\cdots<i_k} p_{i_1} \cdots p_{i_k} \right).$$

Particularly when the parameters are large, numerical evaluation of $\pi(k,N)$ becomes prohibitive. As can be intuitively understood, taking for $\boldsymbol{p} = (p_1, \ldots, p_N)$ the uniform distribution *maximizes* the uniqueness probability, cf. [9,18].

This paper explores birthday problems in the rare-event regime, that is, we focus on situations in which $\pi(k,N)$ is small. The scaling used relies on the parameterization $k \equiv aN$ (where $a \in (0,1)$; assume for ease aN is integer), under which the uniqueness probability decays roughly exponentially. Indeed, applying Stirling's approximation for the factorial, we obtain for the uniform case

$$\pi(aN,N) \sim \frac{1}{\sqrt{1-a}} e^{-Na} (1-a)^{-(1-a)N}$$

(where '$\sim$' indicates that the ratio of the left hand side and right hand side tends to 1 as $N \to \infty$).

Our objective is to quantify the effect of non-uniformity. As argued above, the setup of arbitrary distributions $\boldsymbol{p}$ leads to numerical infeasibilities for larger parameters. We therefore opt for a compromise: the situation of d classes, such that class i has $f_i N$ elements (assumed to be integer), each of which has probability $\alpha_i/N > 0$, where $\sum_{i=1}^d f_i = 1$ and $\sum_{i=1}^d f_i \alpha_i = 1$. In other words: we consider the situation of d heterogeneous classes, where within each of these classes there is uniformity.

Our contributions are the following.

- In the first place, for our model we compute in Section 3 the exponential decay rate

$$\zeta := \lim_{N\to\infty} \frac{1}{N} \log \pi(aN,N), \tag{2}$$

  relying on elementary large-deviation techniques; for $d = 1$ the Stirling approximation above immediately yields

$$\zeta = \zeta_0 := -a - (1-a)\log(1-a). \tag{3}$$

  The solution is in terms of a system of $d-1$ coupled 2nd order polynomial equations (which in general cannot be solved explicitly for $d > 2$). It turns out to be possible though, to decouple this system such that we obtain $d-1$ equations such that each of them contains only a single unknown.

https://doi.org/10.1017/S026996481300034X Published online by Cambridge University Press

- Parameterizing $\alpha_i = 1 + \varphi_i \varepsilon$ (with $\sum_{i=1}^{d} f_i \varphi_i = 0$), we assess to what extent perturbations (i.e., relative to the uniform distribution) affect the decay rate ζ. Writing

$$\zeta = \sum_{k=0}^{\infty} \beta_k \varepsilon^k,$$

  we develop in Section 4 a technique to compute the β_k *explicitly* in terms of the model parameters $\boldsymbol{f}$, $\boldsymbol{\varphi}$, and a. We perform the procedure for $k = 0, 1, 2$; the same approach can be followed for larger k, but then the formulae become less transparent. Evidently $\beta_0 = \zeta_0$. Furthermore, $\beta_1 = 0$ and $\beta_2 < 0$, reflecting the fact that the uniform distribution maximizes the uniqueness probability.
- Section 5 is devoted to techniques to efficiently estimate $\pi(aN, N)$ relying on stochastic simulation. We present a technique that is logarithmically efficient, that is, in which there is subexponential growth (in the scaling parameter N) of the number of runs required to obtain an estimate with a predefined precision; for naïve, direct simulation typically this number grows essentially exponentially.
- The paper concludes in Section 6 with a series of numerical experiments, showing that (i) the proposed simulation approach is fast and accurate, (ii) assuming all outcomes equally likely would lead to estimates for $\pi(k, N)$ that tend to be orders of magnitude off, and (iii) the power-series based approximations work remarkably well.

The literature on birthday problems, dating back to [16], consists of several branches. Quite a few papers consider the probability that c individuals have the same birthday, where c is not necessarily 2 (but usually assuming uniform probabilities); see e.g. [3,7,8,15] to mention a few contributions.

Klotz [11] uses combinatorial techniques to find an efficient numerical scheme to compute the uniqueness probability under *nonuniform* probabilities. Mase [14] finds a recursion (in k) for the $\pi(k, N)$ using Bell polynomials (which is in agreement with the results in [11]), and also gives a number of accurate approximations. The approach followed by Nunnikhoven [17] has some similarity with the techniques used in Section 3 of the present paper; it quantifies the effect of perturbations (relative to the uniform distribution) on the uniqueness probability. The analysis in [2] exploits the relation with random trees.

In general, in the existing literature there is a strong focus on techniques to determine the minimum k such that $\pi(k, N)$ drops below a certain given probability p (in the spirit of the statement: in the setting of the classical birthday problem one needs 23 individuals to make sure that the uniqueness probability drops below 50%). The large-deviations regime, which is the setting of the present paper, has been hardly considered; its relevance is further motivated in the next section. A similar approach has been followed in a related continuous setting, see [10].

2. PRELIMINARIES AND MOTIVATION

In this section we recapitulate a number of known results, and, in passing, motivate our interest in the large-deviations regime. We do so by first considering the uniform case.

Uniform case. For this uniform case, a well-known approximation is

$$\pi_{\mathrm{u}}(k, N) \approx \exp\left(-\frac{1}{2}\frac{k^2}{N}\right);$$

https://doi.org/10.1017/S026996481300034X Published online by Cambridge University Press

this approximation is motivated in example [3]. It is easily seen from (1), in conjunction with 'Stirling' and the standard limit

$$\left(1-\frac{a}{\sqrt{N}}\right)^N e^{a\sqrt{N}} \to e^{-\frac{a^2}{2}},$$

that this result is exact in a specific asymptotic regime [4, p. 83]:

$$\lim_{N\to\infty} \pi_{\rm u}(a\sqrt{N},N) = \exp\left(-\frac{a^2}{2}\right) \tag{4}$$

(where $a\sqrt{N}$ is to be understood as its nearest integer). From these results, we find an approximation for the minimum k such that $\pi(k,N)$ drops below δ, that is, $k_\delta \approx \sqrt{-2N\log\delta}$. If the scaling $k \equiv aN$ is used, rather than the scaling $k \equiv a\sqrt{N}$, Eqs. (2)–(3) entail that

$$\pi_{\rm u}(k,N) \approx e^{-k}\left(1-\frac{k}{N}\right)^{k-N}. \tag{5}$$

An elementary numerical study shows that for small δ, the approximation $k_\delta \approx \sqrt{-2N\log\delta}$ tends to perform poorly; instead, approximations based on equating (5) to δ work remarkably well. This motivates our interest in the large-deviations regime.

Non-uniform case. The counterpart of (4) was found in example [2, Thm. 4]:

$$\lim_{N\to\infty} \pi(a\sqrt{N},N) = \exp\left(-\frac{a^2}{2}\sum_{i=1}^d f_i\alpha_i^2\right);$$

for equal α_i we recover the approximation $\exp(-a^2/2)$. As in the uniform case, this result helps in finding an accurate approximation for k_δ for relatively large δ:

$$k_\delta \approx \left(\sum_{i=1}^d f_i\alpha_i^2\right)^{-\frac{1}{2}} \times \sqrt{-2N\log\delta}. \tag{6}$$

Again this approximation tends to be inaccurate for small δ. The findings of our work enable us to find the counterpart of approximation (5) for the non-uniform case, which also provides us with more accurate approximations for k_δ in the small-δ regime.

We conclude this section with a result that we do not use in the course of our paper, but which may be of independent interest. It is stated for $d=2$, but generalizes straightforwardly. The crucial idea is to condition on the number ℓ samples that is of group 1; the remaining $aN-\ell$ are then of group 2 (where we tacitly assume that $\ell \le fN$ and $aN-\ell \le (1-f)N$). Hence, recognizing the binomial distribution, the probability of all samples being different reads

$$\pi(k,N) = \sum_{\ell=\max\{0,k-(1-f)N\}}^{\min\{k,fN\}} \binom{k}{\ell} (f\alpha_1)^\ell ((1-f)\alpha_2)^{k-\ell}\, \pi_{\rm u}(\ell,fN)\,\pi_{\rm u}(k-\ell,(1-f)N),$$

where $\pi_{\rm u}(\cdot\,,\cdot)$ refers to a uniqueness probability under a uniform distribution. Using the explicit expression for $\pi_{\rm u}(\cdot\,,\cdot)$, this can be rewritten as

$$k!\left(\sum_{\ell=\max\{0,k-(1-f)N\}}^{\min\{k,fN\}} \binom{fN}{\ell}\binom{(1-f)N}{k-\ell}\left(\frac{\alpha_1}{N}\right)^\ell\left(\frac{\alpha_2}{N}\right)^{k-\ell}\right). \tag{7}$$

https://doi.org/10.1017/S026996481300034X Published online by Cambridge University Press

Let X denote the number of red balls when sampling m balls (without replacements) from an urn with M balls, out of which r are red, that is, X has a *hypergeometric* distribution. We define the z-transform of X by $h(z) \equiv h(z \mid M, m, r)$. Then we have proven the following representation for $\pi(k, N)$.

LEMMA 1: *The uniqueness probability for $d = 2$ equals*

$$\pi(k, N) = k! \left(\frac{\alpha_2}{N}\right)^k \times h\left(\frac{\alpha_1}{\alpha_2} \,\middle|\, N, k, fN\right).$$

It is known that the z-transform of the hypergeometric distribution can be expressed in terms of Gaussian hypergeometric series.

3. DECAY RATE

This section determines the decay rate ζ as the solution of a variational problem. We explicitly present the corresponding first order conditions. For the case $d = 2$, we explicitly compute ζ.

3.1. Two Homogeneous Groups

For ease we first consider the case $d = 2$. As before, our starting point is

$$\pi(aN, N) = \sum_{\ell} \binom{aN}{\ell} (f\alpha_1)^{\ell} ((1-f)\alpha_2)^{aN-\ell}\, \pi_{\mathrm{u}}(\ell, fN)\, \pi_{\mathrm{u}}(aN - \ell, (1-f)N), \tag{8}$$

again with $\pi_{\mathrm{u}}(\cdot, \cdot)$ being the uniqueness probability under the uniform distribution. We wish to compute the decay rate of (8).

To this end, we first verify (use 'Stirling'!) that

$$\lim_{N\to\infty} \frac{1}{N} \log \binom{aN}{\ell N} (f\alpha_1)^{\ell N} ((1-f)\alpha_2)^{(a-\ell)N} = -\ell \log\left(\frac{\ell/a}{f\alpha_1}\right) - (a-\ell) \log\left(\frac{1-\ell/a}{(1-f)\alpha_2}\right). \tag{9}$$

It follows from the expression for ζ in the uniform case that

$$\lim_{N\to\infty} \frac{1}{N} \log \pi_{\mathrm{u}}(\ell N, fN) = -(f-\ell) \log\left(\frac{f-\ell}{f}\right) - \ell \tag{10}$$

and

$$\lim_{N\to\infty} \frac{1}{N} \log \pi_{\mathrm{u}}((a-\ell)N, (1-f)N) = -(1-f-a+\ell) \log\left(\frac{1-f-a+\ell}{1-f}\right) - a + \ell. \tag{11}$$

Observe that the probability (8) consists of a number of terms that is roughly linear in N. Also, any sum of positive terms is (i) larger than the maximal term, and (ii) smaller than the maximal term multiplied with the number of terms. The well-known principle of the largest term, in conjunction with the fact that $N^{-1} \log N \to 0$ as $N \to \infty$, then provides us

https://doi.org/10.1017/S026996481300034X Published online by Cambridge University Press

with the decay rate ζ; it equals the sum of (9), (10), and (11), which we denote by $D(\ell)$ in the sequel, when being maximized over $\ell \in I$, where

$$I := [\max\{f + a - 1, 0\}, \min\{f, a\}].$$

Elementary calculus yields that the first order condition reads

$$D'(\ell) = \log\left(\frac{(a-\ell)(f-\ell)}{\ell(1-f-a+\ell)} \times \frac{\alpha_1}{\alpha_2}\right) = 0; \tag{12}$$

it is easily verified that $D(\cdot)$ is concave (on I). Evidently, the first order condition can be rewritten as

$$\xi_1(\ell) := \ell(1-f-a+\ell)\alpha_2 = (a-\ell)(f-\ell)\alpha_1 =: \xi_2(\ell).$$

Observe that in $\max\{f + a - 1, 0\}$ we have that $\xi_1(\cdot)$ is 0 and $\xi_2(\cdot)$ is positive, whereas in $\min\{f, a\}$ it holds that $\xi_2(\cdot)$ is 0 and $\xi_1(\cdot)$ is positive. In other words: the parabolas have a single intersection point in I. We end up with the following result.

THEOREM 1: *In case $d = 2$, the decay rate ζ of $\pi(aN, N)$ equals $D(\ell^\star)$, where $\ell^\star$ is the unique solution in I of*

$$\ell^2(\alpha_2 - \alpha_1) + \ell[(\alpha_1 - \alpha_2)(f + a) + \alpha_2] - \alpha_1 a f = 0. \tag{13}$$

For $\alpha_1 = \alpha_2 = 1$ this result indeed gives the solution of the classical birthday problem, with $\ell^\star = af$.

3.2. Multiple Homogeneous Groups

The above idea can be extended immediately to the case $d > 2$ groups. Again one can condition on the number of samples in each of the groups, which obey a multinomial distribution with parameters aN and $(f_1\alpha_1, \ldots, f_d\alpha_d)$. We obtain the following result; define $\bar{F} := 1 - F$, with $F := \sum_{i=1}^{d-1} f_i$.

THEOREM 2: *In the case $d \in \{2, 3, \ldots, \}$, the decay rate ζ of $\pi(aN, N)$ equals*

$$\begin{aligned}\zeta = -\inf_{\ell_1,\ldots,\ell_{d-1}} \Bigg(& \sum_{i=1}^{d-1} \ell_i \log\left(\frac{\ell_i}{a \times f_i\alpha_i}\right) + \left(a - \sum_{i=1}^{d-1}\ell_i\right)\log\left(\frac{a - \sum_{i=1}^{d-1}\ell_i}{a\left(1 - \sum_{i=1}^{d-1} f_i\alpha_i\right)}\right) \\ & + \sum_{i=1}^{d-1}(f_i - \ell_i)\log\left(\frac{f_i - \ell_i}{f_i}\right) \\ & + \left(1 - \sum_{i=1}^{d-1} f_i - a + \sum_{i=1}^{d-1}\ell_i\right)\log\left(\frac{\bar{F} - a + \sum_{i=1}^{d-1}\ell_i}{\bar{F}}\right) - a\Bigg),\end{aligned}$$

where the infimum is over I, with

$$I := \left\{\boldsymbol{\ell} : 0 \le \ell_i \le \min\{f_i, a_i\} \text{ for } i = 1, \ldots, d-1;\ a - \bar{F} \le \sum_{i=1}^{d-1}\ell_i \le a\right\}.$$

Note that I is nonempty; it contains the vector $\boldsymbol{\ell}$ with $\ell_i := af_i$ (which is optimal if $\alpha_1 = \cdots = \alpha_d = 1$).

https://doi.org/10.1017/S026996481300034X Published online by Cambridge University Press

The corresponding j-th first order condition is, for $j = 1, \ldots, d-1$,

$$\ell_j \left(1 - \sum_{i=1}^{d-1} f_i \alpha_i\right) \left(1 - \sum_{i=1}^{d-1} f_i - a + \sum_{i=1}^{d-1} \ell_i\right) = \left(a - \sum_{i=1}^{d-1} \ell_i\right) (f_j - \ell_j)\alpha_j \left(1 - \sum_{i=1}^{d-1} f_i\right). \tag{14}$$

For given ℓ_i with $i \neq j$, this is a quadratic equation in ℓ_j. The resulting system of coupled quadratic equations does not allow any explicit results for $d > 2$. It can be rewritten as single equations in single unknowns, as follows.

Insert $j = 1$ into Eq. (14), and divide the left hand side and right hand side of the resulting equation by the left hand side and right hand side of Eq. (14). This leads to

$$\ell_j = \frac{f_j \alpha_j \ell_1}{f_1 \alpha_1 - (\alpha_1 - \alpha_j)\ell_1}, \quad \text{and} \quad L := \sum_{j=1}^{d-1} \ell_j = \sum_{j=1}^{d-1} \left(\frac{f_j \alpha_j \ell_1}{f_1 \alpha_1 - (\alpha_1 - \alpha_j)\ell_1}\right).$$

At the same time, (14) implies that

$$L = \frac{a(f_1 - \ell_1)\alpha_1 \bar{F} + (a - \bar{F})\ell_1 \left(1 - \sum_{j=1}^{d-1} f_j \alpha_j\right)}{(f_1 - \ell_1)\alpha_1 \bar{F} + \ell_1 \left(1 - \sum_{j=1}^{d-1} f_j \alpha_j\right)}.$$

Equating both expressions for L leads to a single equation with a single unknown (viz. ℓ_1). Analogous equations can be derived for $i = 2, \ldots, d-1$. We obtain the following result, which reduces to $\boldsymbol{\ell}^\star = a\boldsymbol{f}$ in case $\boldsymbol{\alpha} = \mathbf{1}$.

PROPOSITION 1: *The i-th entry of the vector $\boldsymbol{\ell}^\star$ solves, for $i = 1, \ldots, d-1$,*

$$\sum_{j=1}^{d-1} \left(\frac{f_j \alpha_j \ell_i}{f_i \alpha_i - (\alpha_i - \alpha_j)\ell_i}\right) = \frac{a(f_i - \ell_i)\alpha_i \bar{F} + (a - \bar{F})\ell_i \left(1 - \sum_{j=1}^{d-1} f_j \alpha_j\right)}{(f_i - \ell_i)\alpha_i \bar{F} + \ell_i \left(1 - \sum_{j=1}^{d-1} f_j \alpha_j\right)},$$

and $\ell_d^\star = a - \sum_{i=1}^{d-1} \ell_i^\star$.

4. ANALYSIS OF DECAY RATE

Above we saw that we obtain $\ell_i = a f_i$ if $\alpha_i = 1$ for all i, and that the system does not yield any explicit results if the α_i are distinct. To remedy that last effect, we consider in this section the case of $\alpha_i = 1 + \varphi_i \varepsilon$, in order to see how this perturbation affects the decay rate.

To this end, we write $\zeta = \sum_{k=0}^{\infty} \beta_k \varepsilon^k$ and $\ell_i^\star = \sum_{k=0}^{\infty} \kappa_{ki} \varepsilon^k$, and we wish to determine the coefficients β_k and κ_{ki}, for $i = 1, \ldots, d$ and $k = 0, 1, \ldots$. Interestingly, it turns out that these coefficients can be determined explicitly (in terms of the model parameters $\boldsymbol{f}$, $\boldsymbol{\varphi}$ and a). In this section we demonstrate this procedure for $k = 0, 1, 2$, but it is stressed that the same procedure can be used to determine closed-form expressions for the higher-order coefficients.

First observe that $\kappa_{0i} = a f_i$ and $\beta_0 = \zeta_0 = -a - (1 - a)\log(1 - a)$. In order to obtain β_1 and β_2, we set $\ell_i^\star = a f_i + \kappa_{1i}\varepsilon + \kappa_{2i}\varepsilon^2 + O(\varepsilon^3)$. The j-th first order condition (14) reads,

https://doi.org/10.1017/S026996481300034X Published online by Cambridge University Press

up to $O(\varepsilon^2)$ terms,

$$(af_j+\kappa_{1j}\varepsilon+\kappa_{2j}\varepsilon^2)\left(1-\sum_{i=1}^{d-1}f_i(1+\varphi_i\varepsilon)\right)\left(1-\sum_{i=1}^{d-1}f_i-a+\sum_{i=1}^{d-1}(af_i+\kappa_{1i}\varepsilon+\kappa_{2i}\varepsilon^2)\right)$$
$$=\left(a-\sum_{i=1}^{d-1}(af_i+\kappa_{1i}\varepsilon+\kappa_{2i}\varepsilon^2)\right)(f_j-af_j-\kappa_{1j}\varepsilon-\kappa_{2j}\varepsilon^2)(1+\varphi_j\varepsilon)\left(1-\sum_{i=1}^{d-1}f_i\right),$$

from which the κ_{1i} and κ_{2i} need to be found.

- In order to find $\boldsymbol{\kappa}_1$, we have to solve, recalling that $\bar{F}:=1-F$ with $F:=\sum_{i=1}^{d-1}f_i$,

$$(1-a)\bar{F}\kappa_{1j}-a(1-a)f_j\left(\sum_{i=1}^{d-1}f_i\varphi_i\right)+af_j\sum_{i=1}^{d-1}\kappa_{1i}$$
$$=-a\bar{F}\kappa_{1j}+a(1-a)\bar{F}f_j\varphi_j-(1-a)f_j\sum_{i=1}^{d-1}\kappa_{1i},$$

or

$$\bar{F}\kappa_{1j}+f_j\sum_{i=1}^{d-1}\kappa_{1i}=G_j:=a(1-a)f_j\left(\sum_{i=1}^{d-1}f_i\varphi_i+\bar{F}\varphi_j\right). \tag{15}$$

We are left with a linear system of equations that allows an explicit solution, as is seen as follows. Summing both sides of Eq. (15) over $j=1,\ldots,d-1$, we obtain, with $K_1:=\sum_{k=1}^{d-1}\kappa_{1i}$, the equation $\bar{F}K_1+FK_1=\sum_{j=1}^{d-1}G_j$, so that

$$K_1=\sum_{j=1}^{d-1}G_j=a(1-a)\Phi_f,\quad\text{with}\quad\Phi_f:=\sum_{j=1}^{d-1}f_j\varphi_j.$$

It now immediately follows that

$$\kappa_{1j}=\frac{G_j-f_jK_1}{\bar{F}}=a(1-a)f_j\varphi_j.$$

This equation has an appealing intuitive explanation: if φ_i is positive (negative), then within the sample of aN there are most likely more (less, respectively) items of type i than the usual af_iN. This type of argument gives insight into the behavior of the system, *conditional* on the rare event of uniqueness.

- As we demonstrate now, $\boldsymbol{\kappa}_2$ can be identified in a similar manner; to this end, consider all ε^2-terms. With $K_2:=\sum_{i=1}^{d-1}\kappa_{2i}$, we have to solve

$$\kappa_{2j}\bar{F}^2+K_2f_j\bar{F}=H_j:=f_jK_1\times\left(a\Phi_f-(1-a)\varphi_j\bar{F}\right)+\kappa_{1j}\bar{F}$$
$$\times\left((1-a)\Phi_f-a\varphi_j\bar{F}\right). \tag{16}$$

We determine K_2 by summing Eq. (16) over $j=1,\ldots,d-1$. Defining $\Phi_\kappa:=\sum_{j=1}^{d-1}\kappa_{1j}\varphi_j$, we obtain that

$$K_2=\frac{1}{\bar{F}}\sum_{j=1}^{d-1}H_j=a\left(\frac{F}{\bar{F}}K_1\Phi_f-\bar{F}\Phi_\kappa\right).$$

https://doi.org/10.1017/S026996481300034X Published online by Cambridge University Press

The vector $\boldsymbol{\kappa}_2$ now follows from

$$\kappa_{2j} = \frac{H_j - K_2 f_j \bar{F}}{\bar{F}^2}.$$

Observe the similarities between the procedures to find $\boldsymbol{\kappa}_1$ and $\boldsymbol{\kappa}_2$. Evidently, this approach can be extended to explicitly compute the Taylor approximation for $\boldsymbol{\ell}^\star$ of arbitrary order. To compute $\boldsymbol{\kappa}_k$, though, $\boldsymbol{\kappa}_1$ up to $\boldsymbol{\kappa}_{k-1}$ need to be known (i.e., the procedure is recursive).

To compute the second-order Taylor approximation of the resulting decay rate ζ, we plug $\ell_i^\star = a f_i + \kappa_{1i}\varepsilon + \kappa_{2i}\varepsilon^2 + O(\varepsilon^3)$ into our objective function (see Theorem 2). Elementary Taylor expansions yield that, as $x \to 0$,

$$\log(A + Bx + Cx^2) = \log A + \frac{B}{A}\,x + \left(\frac{C}{A} - \frac{B^2}{2A^2}\right) x^2 + O(x^3),$$

which implies that

$$\begin{aligned}(A + Bx + Cx^2)\log(A + Bx + Cx^2) = A\log A + B(\log A + 1)x \\ + \left(C + \frac{B^2}{2A} + C\log A\right) x^2 + O(x^3).\end{aligned}$$

Likewise

$$\begin{aligned}(A + Bx + Cx^2)\log(D + Ex) = A\log D + \left(B\log D + \frac{AE}{D}\right) x \\ + \left(-\frac{AE^2}{2D^2} + \frac{BE}{D} + C\log D\right) x^2 + O(x^3).\end{aligned}$$

We obtain that $\beta_0 = \zeta_0 = -a - (1-a)\log(1-a)$ (as expected) and $\beta_1 = 0$, so that the decay rate looks like $\beta_0 + \beta_2\varepsilon^2 + O(\varepsilon^3)$. Lengthy computations reveal that $\boldsymbol{\kappa}_2$ cancels entirely; we eventually arrive at

$$\begin{aligned}\beta_2 &= -\sum_{i=1}^{d}\left(\frac{\kappa_{1i}^2}{2af_i} + \frac{\kappa_{1i}^2}{2(1-a)f_i}\right) - \sum_{i=1}^{d-1}\left(\frac{1}{2}af_i\varphi_i^2 - \kappa_{1i}\varphi_i\right) - \frac{1}{2}a\frac{\Phi_f^2}{\bar{F}} + K_1\frac{\Phi_f}{\bar{F}} \\ &= -\sum_{i=1}^{d}\left(\frac{\kappa_{1i}^2}{2af_i} + \frac{\kappa_{1i}^2}{2(1-a)f_i}\right) - \sum_{i=1}^{d}\left(\frac{1}{2}af_i\varphi_i^2 - \kappa_{1i}\varphi_i\right) \\ &= -\left(\sum_{i=1}^{d} f_i\varphi_i^2\right) \times \frac{a^2}{2} < 0.\end{aligned}$$

The negative sign of β_2 confirms that any deviation from the uniform distribution *reduces* the probability of uniqueness. The nest theorem summarizes our findings.

THEOREM 3: *Suppose* $\alpha_i = 1 + \varphi_i\varepsilon$, *for* $i = 1, \ldots, d$. *Then, as* $\varepsilon \to 0$,

$$\zeta = \zeta_0 - \left(\sum_{i=1}^{d} f_i(\varphi_i\varepsilon)^2\right) \times \frac{a^2}{2} + O(\varepsilon^3).$$

https://doi.org/10.1017/S026996481300034X Published online by Cambridge University Press

Remark 1: The above result can easily be verified for the special case $a = 1$. First observe that in this case

$$\pi(N,N) = \binom{N}{f_1 N, \ldots, f_d N} \prod_{i=1}^{d} \left((\alpha_i f_i)^{f_i N} \pi_{\mathrm{u}}(f_i N, f_i N) \right).$$

Applying that $\pi_{\mathrm{u}}(N,N) = N!/N^N \approx \sqrt{2\pi N} e^{-N}$, we obtain in the usual way that $\zeta = -1 + \sum_{i=1}^{d} f_i \log \alpha_i$. Now plugging in $\alpha_i = 1 + \varphi_i \varepsilon$, we obtain, up to $O(\varepsilon^3)$ terms,

$$\zeta = -1 + \sum_{i=1}^{d} f_i \log(1 + \varphi_i \varepsilon) = -1 + \sum_{i=1}^{d} f_i \left(\varphi_i \varepsilon - \frac{1}{2} \varphi_i^2 \varepsilon^2 \right) = -1 - \frac{1}{2} \sum_{i=1}^{d} f_i (\varphi_i \varepsilon)^2.$$

This last expression is in agreement with Theorem 3 for $a = 1$.

Remark 2: Realizing that $\varphi_i \varepsilon = \alpha_i - 1$, Theorem 3 leads to the approximation

$$\pi(aN,N) \approx \pi_\varepsilon(aN,N) := \exp\left(N \left(\zeta_0 - \frac{a^2}{2} \left(\sum_{i=1}^{d} f_i \alpha_i^2 - 1 \right) \right) \right); \tag{17}$$

here it is used that $\sum_{i=1}^{d} f_i \varphi_i^2 = \sum_{i=1}^{d} f_i \alpha_i^2 - 1$, due to $\sum_{i=1}^{d} f_i = \sum_{i=1}^{d} f_i \alpha_i = 1$.

Remark 3: The Kullback–Leibler divergence of the nonuniform probabilities $(1 + \varphi_i \varepsilon)/N$ with respect to the uniform probabilities $1/N$ reads

$$\sum_{i=1}^{d} f_i N \left(\frac{1 + \varphi_i \varepsilon}{N} \right) \log \left(\left(\frac{1 + \varphi_i \varepsilon}{N} \right) \Big/ \left(\frac{1}{N} \right) \right) = \frac{1}{2} \sum_{i=1}^{d} f_i (\varphi_i \varepsilon)^2 + O(\varepsilon^3).$$

Comparing this with Theorem 3, we conclude that apparently the Kullback–Leibler divergence indicates how the decay rate ζ_0 (based on the uniform distribution) needs to be adapted to appropriately take into account the effect of the non-uniformity.

5. IMPORTANCE SAMPLING SIMULATIONS

As observed above, we have the representation

$$\pi(k,N) = \sum_{\boldsymbol{\ell}} \binom{N}{\ell_1, \ldots, \ell_d} \prod_{i=1}^{d} (f \alpha_i)^{\ell_i} \pi_{\mathrm{u}}(\ell_i, f_i N), \tag{18}$$

where the summation is over all $\ell_i \in \{0, \ldots, f_i N\}$ such that $\sum_{i=1}^{d} \ell_i = k$. Direct numerical evaluation is problematic, particularly when N is large (although substantial improvements can be made using the techniques developed in [11]). One option could be to resort to stochastic simulation. In this section, we first present a straightforward method, and argue that this naïve approach tends to be slow. Then we show that this problem can be overcome relying on importance sampling techniques. We present these findings for $d = 2$, but the conclusions hold for any d.

https://doi.org/10.1017/S026996481300034X Published online by Cambridge University Press

5.1. Naïve Approach

Take, as mentioned above, $d = 2$, and suppose we wish to estimate $\pi(aN, N)$. Define

$$X_i \equiv X_i(N) := \pi_{\mathrm{u}}(L_i, fN)\pi_{\mathrm{u}}(aN - L_i, (1-f)N),$$

with the $L_i \equiv L_i(N)$ i.i.d. binomial random variables with parameters aN and $f\alpha_1$. Then an unbiased estimator of the probability under study is

$$\hat{\pi}(aN, N) := \frac{1}{n}\sum_{i=1}^{n} X_i(N).$$

It is clear that

$$\lim_{N\to\infty} \frac{1}{N} \log \mathbb{E}\left(\left(\frac{1}{n}\sum_{i=1}^{n} X_i(N)\right)^2\right) \geq \lim_{N\to\infty} \frac{2}{N} \log \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i(N)\right) = 2\zeta. \tag{19}$$

In the context of this setup, we call a simulation technique *logarithmically efficient* if this lower bound is actually achieved, that is,

$$\lim_{N\to\infty} \frac{1}{N} \log \mathbb{E}\left(\left(\frac{1}{n}\sum_{i=1}^{n} X_i(N)\right)^2\right) = 2\zeta;$$

one could say that, in view of (19), on a logarithmic scale, such a simulation procedure has optimal performance. In our situation, it essentially means that the number of runs required to obtain an estimate with a given precision (defined as the confidence interval width relative to the estimate) and confidence grows subexponentially in N, see example [1].

First observe that

$$\lim_{N\to\infty} \frac{1}{N} \log \mathbb{E}\left(\left(\frac{1}{n}\sum_{i=1}^{n} X_i(N)\right)^2\right) = \lim_{N\to\infty} \frac{1}{N} \log \mathbb{E}\left(X(N)^2\right), \tag{20}$$

with $X(N) := \pi_{\mathrm{u}}(L, fN)\pi_{\mathrm{u}}(aN - L, (1-f)N)$, and $L \equiv L(N)$ being a binomial random variable with parameters aN and $f\alpha_1$. Now, as before,

$$\mathbb{E}\left(X(N)^2\right) = \sum_{\ell} \binom{aN}{\ell} (f\alpha_1)^{\ell}((1-f)\alpha_2)^{aN-\ell} \times \pi_{\mathrm{u}}^2(\ell, fN)\, \pi_{\mathrm{u}}^2(aN - \ell, (1-f)N),$$

such that decay rate (20) reads

$$-\inf_{\ell\in I}\left(\ell \log\left(\frac{\ell/a}{f\alpha_1}\right) + (a-\ell)\log\left(\frac{1-\ell/a}{(1-f)\alpha_2}\right)\right.$$
$$\left.\times\ 2(f-\ell)\log\left(\frac{f-\ell}{f}\right) + 2(1-f-a+\ell)\log\left(\frac{1-f-a+\ell}{1-f}\right) - 2a\right).$$

It turns out that this implies exponential growth of the number of runs required to obtain an estimate of a given precision and confidence; this number is not inversely proportional to the probability to be estimated though (unlike in many other rare event simulation examples).

https://doi.org/10.1017/S026996481300034X Published online by Cambridge University Press

5.2. Importance Sampling

In this case, we sample, in an i.i.d. manner, the L_i again from a binomial distribution, but now with parameters aN and $\ell^\star/a$, with $\ell^\star$ solving (13). To obtain an unbiased estimator, the resulting X_i have to be multiplied with appropriate likelihood ratios. The idea is that, variance reduction can be achieved by sampling under an alternative measure.

We denote

$$Y_i \equiv Y_i(N) := \left(\frac{f\alpha_1}{\ell^\star/a}\right)^{L_i} \left(\frac{(1-f)\alpha_2}{1-\ell^\star/a}\right)^{aN-L_i} \pi_{\rm u}(L_i, fN)\,\pi_{\rm u}(aN-L_i,(1-f)N),$$

and construct an unbiased estimator through

$$\hat{\pi}_{\rm IS}(aN,N) := \frac{1}{n}\sum_{i=1}^{n} Y_i(N);$$

the idea here is that this estimator is unbiased due to the fact that the individual 'observations', that is, the $\pi_{\rm u}(L_i, fN)\,\pi_{\rm u}(aN-L_i,(1-f)N)$ for $i=1,\ldots,n$, are weighted by *likelihood ratios*

$$\left(\frac{f\alpha_1}{\ell^\star/a}\right)^{L_i} \left(\frac{(1-f)\alpha_2}{1-\ell^\star/a}\right)^{aN-L_i}$$

that reflect the difference in probability between the original measure (to be associated with the binomial distribution with parameters aN and $\alpha_1 f$) and the alternative measure (that relates to the binomial distribution with parameters aN and $\ell^\star/a$).

To verify logarithmic efficiency of this procedure, we have to identify the decay rate of the second moment of $Y(N)$, that is, the decay rate of

$$\sum_{\ell} \binom{aN}{\ell} (f\alpha_1)^{\ell}((1-f)\alpha_2)^{aN-\ell} \left(\frac{f\alpha_1}{\ell^\star/a}\right)^{\ell} \left(\frac{(1-f)\alpha_2}{1-\ell^\star/a}\right)^{aN-\ell} \times \pi_{\rm u}^2(\ell, fN)\,\pi_{\rm u}^2(aN-\ell,(1-f)N).$$

Using the same approach as in Section 3.1, we obtain that this decay rate equals

$$-\inf_{\ell\in I}\left(\ell\log\left(\frac{\ell/a\times\ell^\star/a}{(f\alpha_1)^2}\right)+(a-\ell)\log\left(\frac{(1-\ell/a)(1-\ell^\star/a)}{((1-f)\alpha_2)^2}\right) + 2(f-\ell)\log\left(\frac{f-\ell}{f}\right)+2(1-f-a+\ell)\log\left(\frac{1-f-a+\ell}{1-f}\right)-a\right).$$

The corresponding first order condition reads

$$\log\left(\frac{(a-\ell)(a-\ell^\star)(f-\ell)^2}{\ell\,\ell^\star(1-f-a+\ell)^2}\times\left(\frac{\alpha_1}{\alpha_2}\right)^2\right)=0,$$

which is solved for $\ell=\ell^\star$, cf. (12). This directly implies that the decay rate under consideration equals $2D(\ell^\star)=2\zeta$, which implies logarithmic efficiency.

THEOREM 4: *The estimator $\hat{\pi}_{\rm IS}(aN,N)$ is logarithmically efficient.*

https://doi.org/10.1017/S026996481300034X Published online by Cambridge University Press

6. NUMERICAL EXPERIMENTS

In this section, we present two types of experiments: (1) experiments that quantify the effect of the non-uniformity, and (2) experiments that assess the performance of the approximations of Section 5 (e.g., the approximation highlighted in Remark 2).

6.1. The Impact of Non-uniformity

Suppose a questionnaire is filled out by 120 people, and suppose one has to reveal the year and month in which he or she was born. Neglecting people older than 100 years, this means that the number of possible outcomes is $100 \times 12 = 1200$. Notice, however, that these 1200 outcomes are generally *not* equally likely. The current population pyramid of the city of Amsterdam, for instance, shows a peak around 1980 with about 2% of the population was born that year, whereas just about 8% of the population was born before 1940. In previous studies, we have developed techniques to assess the fraction of uniquely identifiable individuals, see [12,13]. Our objective here is to determine the probability that within a sampled group of 120 persons, everyone has a unique year/month combination.

A procedure that could be followed is the following. Let γ_i be the relative frequency of year/month combination i, for $i = 1, \ldots, 1200$. First order the year/month combinations in decreasing order of relative frequency. Then we cluster them into, say, 10 groups of size 120 each, and set

$$\alpha_i := \frac{1}{120} \sum_{j=120(i-1)+1}^{120\,i} \gamma_j;$$

in the notation of our paper, we have $f_i = 0.1$, for $i = 1, \ldots, 10$. (If the relative frequencies are still quite variable within the groups of size 120, then one could opt for splitting the 1200 combinations in more than 10 groups.) In the language of our paper, we now wish to determine $\pi(120, 1, 200)$.

It is first observed that direct computation through representation (18) is often computationally infeasible. Realize that

$$\#\left\{\boldsymbol{\ell} : \ell_i \in \{0, \ldots, M\}, \sum_{i=1}^{d} \ell_i = M\right\} \approx \frac{1}{d-1} M^{d-1},$$

so that in our example the summation (18) would consist of roughly $9^{-1} \times 120^9 \approx 5.7 \times 10^{17}$ terms (take $M = 120$ and $d = 10$), which is evidently prohibitive from a numerical standpoint. We therefore resort to simulation.

To evaluate the uniqueness probability in the above scenario, we take the model described in this paper, with $N = 1200$ and $a = 0.1$. In the example of the population of Amsterdam, it turns out that we (by approximation) have $\boldsymbol{\alpha} = (2.0, 1.8, 1.6, 1.3, 1.0, 0.8, 0.6, 0.4, 0.3, 0.2)'$; it is readily verified that indeed $\sum_{i=1}^{10} f_i \alpha_i = 1$. It is a matter of straightforward numerical analysis to obtain that

$$\ell^\star / a = (0.190, 0.172, 0.154, 0.129, 0.103, 0.085, 0.065, 0.045, 0.034, 0.023)'.$$

Comparing the entries of this vector with the $\alpha_i f_i$, we observe that the importance of sampling is set up such that (roughly) outcomes in the groups i with $\alpha_i > 1$ are getting less likely (compared to the naïve simulation, that is, sampling under the original measure), whereas outcomes in the groups i with $\alpha_i < 1$ are getting more likely (as expected).

https://doi.org/10.1017/S026996481300034X Published online by Cambridge University Press

TABLE 1. Uniqueness probabilities; a varying (left panel) and N varying (right panel)

a	$\hat{\pi}(aN,N)$	$\pi_{\rm u}(aN,N)$	$e^{\zeta_0 N}$	N	$\hat{\pi}(aN,N)$	$N^{-1}\times\log\hat{\pi}(aN,N)$	$\pi_{\rm u}(aN,N)$
0.02	7.26×10^{-1}	7.93×10^{-1}	7.85×10^{-1}	200	2.58×10^{-1}	-6.77×10^{-3}	3.74×10^{-1}
0.04	2.69×10^{-1}	3.85×10^{-1}	3.73×10^{-1}	400	6.21×10^{-2}	-6.94×10^{-3}	1.33×10^{-1}
0.06	5.02×10^{-2}	1.14×10^{-1}	1.10×10^{-1}	600	1.50×10^{-2}	-6.99×10^{-3}	4.72×10^{-2}
0.08	4.64×10^{-3}	2.01×10^{-2}	1.93×10^{-2}	800	3.61×10^{-3}	-7.03×10^{-3}	1.67×10^{-2}
0.10	2.09×10^{-4}	2.12×10^{-3}	2.01×10^{-3}	1000	8.70×10^{-4}	-7.05×10^{-3}	5.96×10^{-3}
0.12	4.58×10^{-6}	1.30×10^{-4}	1.22×10^{-4}	1200	2.09×10^{-4}	-7.06×10^{-3}	2.12×10^{-3}
0.14	4.72×10^{-8}	4.66×10^{-6}	4.33×10^{-6}	1400	5.05×10^{-5}	-7.07×10^{-3}	7.52×10^{-4}
0.16	2.28×10^{-10}	9.54×10^{-8}	8.74×10^{-8}	1600	1.22×10^{-5}	-7.07×10^{-3}	2.67×10^{-4}
0.18	5.07×10^{-13}	1.10×10^{-9}	9.99×10^{-9}	1800	2.92×10^{-6}	-7.08×10^{-3}	9.48×10^{-5}
0.20	5.08×10^{-16}	7.10×10^{-12}	6.35×10^{-12}	2000	7.07×10^{-7}	-7.08×10^{-3}	3.37×10^{-5}

- *Experiment I:* In the left panel of Table 1 we present the uniqueness probabilities corresponding to the example above, but for various values of a, and $N=1200$. The second column (i.e., $\hat{\pi}(aN,N)$) gives estimates for $\pi(aN,N)$ using the efficient simulation techniques proposed in Section 5—the time needed to generate one entry in this column, with 95% confidence and 10% precision, is well below 1 second on a standard PC. In the third column (i.e., $\pi_{\rm u}(aN,N)$) we do as if the 1200 outcomes were uniformly distributed; as expected, we observe that this would lead to a gross overestimation of the probability under consideration (particularly for somewhat larger values of a). The last column shows that the closed-form expression

$$e^{-\zeta_0 N}=\exp\left(-aN-(1-a)\log(1-a)N\right)$$

  is an accurate approximation for $\pi_{\rm u}(aN,N)$.
- *Experiment II:* We use the same $\boldsymbol{f}$ and $\boldsymbol{\alpha}$ as in the first experiment, but now we fix a at 0.1, and vary N. In the right panel of Table 1 the second column (i.e., $\hat{\pi}(aN,N)$) displays estimates for $\pi(aN,N)$ using the efficient simulation techniques of Section 5. Then we compute in the third column $N^{-1}\times\log\hat{\pi}(aN,N)$; it is seen that $N^{-1}\times\log\hat{\pi}(aN,N)$ converges to $\zeta=-7.08\times10^{-3}$ as $N\to\infty$, indicating that $\pi(aN,N)$ decays exponentially; this convergence is quite rapid. The last column gives the uniqueness probabilities if all outcomes for the situation that all outcomes are equally likely, and we again observe that this naïve estimate can be orders of magnitude off.
- *Experiment III:* Above we described how we split the population into ten more or less uniform groups. One may wonder how this stratification affects the results. To study this effect, we looked at a coarser stratification, and check to what extent the probabilities change. To this end, we take $d=5$ and $\boldsymbol{\alpha}=(1.9,1.45,0.9,0.5,0.25)'$, with the f_i (for $i=1,\ldots,5$) being equal to 0.2. The results are found in Table 2. We observe that there is a modest deviation with respect to the case of $d=10$ (towards the $\pi_{\rm u}(aN,N)$). Additional experiments (not reported here) indicate that further coarsening (for instance into three groups) leads to more substantial deviations with respect to the case of $d=10$.

https://doi.org/10.1017/S026996481300034X Published online by Cambridge University Press

TABLE 2. Uniqueness probabilities for coarser grouping; a varying (left panel) and N varying (right panel).

a	$\hat{\pi}(aN, N)$	N	$\hat{\pi}(aN, N)$	$N^{-1} \times \log \hat{\pi}(aN, N)$
0.02	7.28×10^{-1}	200	2.61×10^{-1}	-6.71×10^{-3}
0.04	2.72×10^{-1}	400	6.38×10^{-2}	-6.88×10^{-3}
0.06	5.15×10^{-2}	600	1.55×10^{-2}	-6.94×10^{-3}
0.08	4.84×10^{-3}	800	3.79×10^{-3}	-6.97×10^{-3}
0.10	2.24×10^{-4}	1000	9.19×10^{-4}	-6.99×10^{-3}
0.12	5.04×10^{-6}	1200	2.24×10^{-4}	-7.00×10^{-3}
0.14	5.40×10^{-8}	1400	5.47×10^{-5}	-7.01×10^{-3}
0.16	2.70×10^{-10}	1600	1.33×10^{-5}	-7.01×10^{-3}
0.18	6.38×10^{-13}	1800	3.24×10^{-6}	-7.02×10^{-3}
0.20	6.82×10^{-16}	2000	7.91×10^{-7}	-7.02×10^{-3}

TABLE 3. Taylor approximation for $\boldsymbol{\ell}^\star$.

	1	2	3	4	5	6	7	8	9	10
$\boldsymbol{\ell}^\star$	0.0179	0.0164	0.0147	0.0130	0.0112	0.0094	0.0075	0.0054	0.0033	0.0011
$a\boldsymbol{f} + \boldsymbol{\kappa}_1\varepsilon$	0.0181	0.0163	0.0145	0.0127	0.0109	0.0091	0.0073	0.0055	0.0037	0.0019

TABLE 4. Various approximations for the uniqueness probabilities; a varying.

a	$\hat{\pi}(aN, N)$	$\pi_{\mathrm{u}}(aN, N)$	$\exp(-\zeta_0 N)$	$\pi_\varepsilon(aN, N)$
0.05	1.90×10^{-1}	2.88×10^{-1}	2.80×10^{-1}	1.86×10^{-1}
0.10	1.09×10^{-3}	5.96×10^{-3}	5.65×10^{-3}	1.09×10^{-3}
0.15	1.58×10^{-7}	7.67×10^{-6}	7.07×10^{-6}	1.73×10^{-7}
0.20	4.55×10^{-13}	5.21×10^{-10}	4.66×10^{-10}	6.34×10^{-13}

6.2. Assessment of the Approximations of the Decay Rate

In the second set of experiments we assess the accuracy of the estimates of Section 4. As before we take $d = 10$ and $f_i = 0.1$ for $i = 1, \ldots, 10$, but we now choose $\boldsymbol{\alpha} = (1.9, 1.7, 1.5, 1.3, 1.1, 0.9, 0.7, 0.5, 0.3, 0.1)'$.

- *Experiment IV:* In this experiment we pick $N = 1000$. In Table 3 we present, for $a = 0.1$, the optimizer $\ell^\star$ with its first order Taylor approximation; we conclude that they match well. In Table 4 we compare, for various values of a, $\hat{\pi}(aN, N)$, $\pi_{\mathrm{u}}(aN, N)$, $\exp(\zeta_0 N)$, and $\pi_\varepsilon(aN, N)$ (as given in Remark 2). The table shows that adding the ε^2-term improves the accuracy of the approximation considerably.
- *Experiment V:* We now consider the minimum number k_δ that ensures that the uniqueness probability is below a probability δ. We compare in Table 5 (i) the 'exact' outcome, obtained by trial and error (where the uniqueness probabilities are computed using our simulation procedure), (ii) Formula (6) (denoted by $k_\delta^{(1)}$), (iii) by equating (17) to δ and solving $k \equiv aN$ (denoted by $k_\delta^{(2)}$). It is observed that estimating k_δ relying on the approximation (17) results in a substantial improvement over (6), particularly for small δ.

https://doi.org/10.1017/S026996481300034X Published online by Cambridge University Press

TABLE 5. Values for k_δ; $N = 100$ (left panel) and $N = 1000$ (right panel).

δ	k_δ	$k_\delta^{(1)}$	$k_\delta^{(2)}$	δ	k_δ	$k_\delta^{(1)}$	$k_\delta^{(2)}$
10^{-2}	26	26	26	10^{-2}	83	83	83
10^{-4}	35	35	37	10^{-4}	116	116	117
10^{-6}	43	43	46	10^{-6}	141	142	144
10^{-8}	48	49	52	10^{-8}	162	163	166
10^{-10}	53	54	59	10^{-10}	181	182	185
10^{-12}	58	59	64	10^{-12}	197	198	204
10^{-14}	61	63	70	10^{-14}	212	214	220

7. DISCUSSION AND OUTLOOK

This paper considered uniqueness probabilities in generalized birthday problems, with a focus on the large deviations regime. Scaling the sample size proportional to the number of possible outcomes, the uniqueness probability decays roughly exponentially. We determined the associated decay rate (in terms of a variational problem), developed (explicit) accurate approximations for this decay rate, devised efficient simulation techniques, and performed a broad set of numerical experiments that offered additional insights.

Several challenging extensions can be thought of. One could for instance of the *dynamic birthday problem*, where individuals arrive according to a Poisson process (with rate λ) and leave after some random time (where these sojourn times are assumed i.i.d. with mean μ). The (stationary) uniqueness probability in this model can be regarded as the uniqueness probability in a model in which the size of the sample is random (i.e., Poisson distributed with mean $\nu := \lambda/\mu$; recognize the M/G/∞ queue), and is relatively straightforward to compute. Indeed, in the uniform case we obtain

$$\sum_{k=0}^{N} e^{-\nu} \frac{\nu^k}{k!} \pi_{\rm u}(k,N) = e^{-\nu} \left(1 + \frac{\nu}{N}\right)^N.$$

When scaling ν by N (i.e., replace ν by νN in the above formula, there is exponential decay in N. Although these stationary metrics are straightforward (also in the non-uniform case), transient metrics are more challenging; think for instance of the probabilistic properties of the time elapsed between two "non-uniqueness events".

In a variant of this model, individuals are only accepted when their birthday does not match with one of the birthdays of the current population. Again various (steady-state and transient) performance metrics can be thought of.

Acknowledgements

The problem was brought to the author's attention by M.R. Koot and C.T.A.M. de Laat (Institute of Informatics, University of Amsterdam). The author also thanks A.V. Gnedin (Utrecht University, the Netherlands) for inspiring discussions.

References

1. Asmussen, S. & Glynn, P. (2007). *Stochastic Simulation: Algorithms and Analysis.* NY: Springer-Verlag.
2. Camarri, M. & Pitman, J. (2000). Limit distributions and random trees derived from the birthday problem with unequal probabilities. *Electronic Journal of Probability* 5: 1–18.
3. Diaconis, P. & Mosteller, F. (1989). Methods for studying coincidences. *Journal of the American Statistical Association* 84: 853–861.

https://doi.org/10.1017/S026996481300034X Published online by Cambridge University Press

4. Durrett, R. (1995). *Probability: Theory & Examples.* Wadsworth-Brooks/Cole, CA: Pacific Grove.
5. Feller, W. (1968). *An Introduction to Probability Theory and its Applications*, Vol. 1, 3rd edn, NY: Wiley.
6. Gail, M., Weiss, G., Mantel, N. & O'Brien, S. (1979). A solution to the generalized birthday problem with application to allozyme screening for cell culture contamination. *Journal of Applied Probability* 16: 242–251.
7. Henze, N. (1998). A Poisson limit law for a generalized birthday problem. *Statistics and Probability Letters* 39: 333–336.
8. Holst, L. (1986). On birthday, collectors', occupancy and other classical urn problems. *International Statistical Review* 54: 15–27.
9. Joag-Dev, K. & Proschan, F. (1992). The birthday problem with unlike probabilities. *American Mathematical Monthly* 99: 10–12.
10. Juneja, S. & Mandjes, M. Overlap problems on the circle. *Journal of Applied Probability* to appear.
11. Klotz, J. (1979). The birthday problem with unequal probabilities. *Technical Report*, No. 59, Department of Statistics, University of Wisconsin.
12. Koot, M., van't Noordende, G., Mandjes, M. & de Laat, C. (2013). A probabilistic perspective on re-identifiability. *Mathematical Population Studies*, 20: 155–171.
13. Koot, M. & Mandjes, M. (2012). The analysis of singletons in generalized birthday problems. *Probability in the Engineering and Informational Sciences* 26: 245–262.
14. Mase, S. (1992). Approximations to the birthday problem with unequal occurrence probabilities and their application to the surname problem in Japan. *Annals of the Institute of Statistical Mathematics* 44: 479–499.
15. McKinney, E. (1966). Generalized birthday problem. *American Mathematical Monthly* 73: 385–387.
16. von Mises, R. (1938). Über Aufteilungs- und Besetzungswahrscheinlichkeiten. *Revue de la Faculté des Sciences de L'Université d'Istanbul* 4: 145–163.
17. Nunnikhoven, T. (1992). A birthday problem solution for non-uniform birth frequencies. *The American Statistician* 46: 270–274.
18. Rust, P. (1976). The effect of leap years and seasonal trends on the birthday problem. *The American Statistician* 30: 197–198.
19. Wagner, D. (2002). A generalized birthday problem. In: Proc. Crypto 2002 *Lecture Notes in Computer Science* Vol. 2442, NY: Springer-Verlag, pp. 288–303.

https://doi.org/10.1017/S026996481300034X Published online by Cambridge University Press