Probability in the Engineering and Informational Sciences, **26**, 2012, 245–262.
doi:10.1017/S0269964811000350

THE ANALYSIS OF SINGLETONS IN GENERALIZED BIRTHDAY PROBLEMS

MATTHIJS R. KOOT
Informatics Institute, University of Amsterdam, The Netherlands
E-mail: koot@cyberwar.nl

MICHEL MANDJES
Korteweg-de Vries Institute for Mathematics,
University of Amsterdam, The Netherlands;
Eurandom, Eindhoven University of Technology, The Netherlands
CWI, Amsterdam, The Netherlands
E-mail: m.r.h.mandjes@uva.nl

This paper describes techniques to characterize the number of singletons in the setting of the generalized birthday problem, that is, the birthday problem in which the birthdays are non-uniformly distributed over the year. Approximations for the mean and variance presented which explicitly indicate the impact of the heterogeneity (expressed in terms of the Kullback–Leibler distance with respect to the homogeneous distribution). Then an iterative scheme is presented for determining the distribution of the number of singletons. The approximations are validated by experiments with demographic data.

1. INTRODUCTION

Consider a population of k people, where each of them is independently assigned a certain 'feature' (for instance: gender, birthday, age, etc.), which is element of $\{1, \ldots, N\}$; in the case of gender obviously $N = 2$, in the case of birthday $N = 365$ (neglecting leap years), in the case of age N can be taken, say, 95. We assume the distribution of the feature over $\{1, \ldots, N\}$ is given, which is not *a priori* assumed to be uniform (birthday and gender will be roughly uniform, whereas age will not). In the literature, this setting is often referred to as that of the *generalized birthday problem*, or, alternatively, the birthday problem with unequal probabilities. There is a vast literature on this topic, e.g. [3–6,9,10].

© Cambridge University Press 2012 0269-9648/12 $25.00

https://doi.org/10.1017/S0269964811000350 Published online by Cambridge University Press

Some of the outcomes will be assigned to just one of the k people in the population; we call these *singletons*. The objective of this paper is to analyse the distribution of the number of singletons S. We subsequently address its mean and variance, as well as a computational scheme for evaluating the distribution of S. It is noted that the previous literature primarily focused on the probability that all k samples are unique (where it was obviously assumed that $k \leq N$).

Our research was motivated by problems arising in privacy and security. Suppose people are asked to fill out a questionnaire in which they have to reveal their zip code and age. One may then wonder to what extent it is possible to uniquely identify these people. More specifically, supposing that k individuals share a zip code, the question reduces to: how many of these k people have a *unique* age? Observe that we are in the setting of the generalized birthday problem described above, and that we are interested in characterizing the distribution of the number of singletons. The lower the number of singletons, the higher the privacy level.

In more detail, our contributions are as follows. Our results cover both the homogeneous (that is, all outcomes $1,\ldots,N$ being equally likely, that is, have probability $1/N$) and heterogeneous settings. In the latter, we assume there are F_i 'bins' that have probability α_i/N; obviously we require that $F_1 + \cdots + F_d = N$ and $\alpha_1 F_1 + \cdots + \alpha_d F_d = N$.

- In Section 2, we first derive an explicit expression for the mean number of singletons $\mathbb{E}S$. We then scale the number of samples and number of outcomes per group by N, that is, $k \equiv aN$ and $F_i \equiv f_i N$. We then show that the mean number of singletons in the scaled model, that is, $\mathbb{E}S(N)$ can be accurately approximated by

$$\mathbb{E}S(N) \approx aN\, e^{-a}\left(1 + \frac{a}{2}(a-2)\kappa\right),$$

  where κ is the Kullback–Leibler distance [2,8] between our heterogeneous distribution and the homogeneous one. This approximation nicely reflects the impact of heterogeneity on the number of singletons. As we will argue, this effect is both quantitatively and qualitatively different for different values of a: for low values of a, $\mathbb{E}S(N)$ is increasing in κ, whereas for high values of a, $\mathbb{E}S(N)$ is decreasing in κ. We illustrate the theory by an example.
- In Section 3, we perform a similar analysis for the variance of S. Again we first derive an exact expression for $\mathbb{V}\mathrm{ar}S$, and then consider approximations in the scaled model.
- Section 4 first develops a recursive algorithm that identifies the full distribution of S for the homogeneous case. A crucial role is played by a technique to find the probability of no singletons, i.e., $\mathbb{P}(S = 0)$. Then it is demonstrated how to extend the analysis to the heterogeneous case, for which also a more explicit approximation is presented.

https://doi.org/10.1017/S0269964811000350 Published online by Cambridge University Press

- Section 5 is devoted to numerical experiments. Based on demographic data of all Dutch municipalities, we estimate the heterogeneity κ, and then assess the accuracy of the approximations for $\mathbb{E}S$ and $\mathbb{V}\mathrm{ar}S$.

2. MEAN NUMBER OF IDENTIFIABLE OBJECTS

In this section, we analyze the mean number of singletons. We find an exact expression, as well as approximations that show how the heterogeneity affects this quantity.

2.1. Explicit Expressions

We first consider the homogeneous case: suppose one throws k balls into N bins, uniformly at random. Then the probability that a given bin contains exactly one ball (a 'singleton') is

$$k \cdot \frac{1}{N}\left(1-\frac{1}{N}\right)^{k-1}, \tag{1}$$

Here we make use of the fact that the number of balls in that bin obeys a binomial distribution with parameter k and $1/N$. As there are N bins, it follows that the mean number of singletons is

$$\mathbb{E}S = k\left(1-\frac{1}{N}\right)^{k-1}.$$

The result for the homogeneous case is standard, but interestingly it can be extended to the heterogeneous setting in a straightforward manner. Let there be F_i bins that have probability α_i/N; obviously $F_1+\cdots+F_d = N$ and $\alpha_1 F_1+\cdots+\alpha_d F_d = N$. Let N_i be the number of balls that end up in bins of type i, and let S_i be the number of singletons among them; observe that N_i has a binomial distribution with parameters k and $\alpha_i F_i/N$. It is clear that

$$\mathbb{E}S_i = kF_i\left(1-\frac{\alpha_i}{N}\right)^{k-1}\frac{\alpha_i}{N}. \tag{2}$$

We arrive at the following statement.

PROPOSITION 1: *In the heterogeneous model defined above, the mean number of singletons equals*

$$\mathbb{E}S = k\sum_{i=1}^{d} F_i\left(1-\frac{\alpha_i}{N}\right)^{k-1}\frac{\alpha_i}{N}.$$

We now consider the number of singletons $S(N)$ in the asymptotic regime in which there are aN balls, and F_i is scaled by N (that is, $F_i \equiv f_i N$). After straightforward calculus we find the following result.

https://doi.org/10.1017/S0269964811000350 Published online by Cambridge University Press

PROPOSITION 2: *In the scaled heterogeneous model defined above, the mean number of singletons satisfies, as* $N \to \infty$,

$$\frac{\mathbb{E}S(N)}{N} \to a \sum_{i=1}^{d} \alpha_i f_i e^{-\alpha_i a}.$$

The above result essentially states that the number of singletons equals roughly Na (the number of balls), but thinned by a factor $\sum_{i=1}^{d} \alpha_i f_i \exp(-\alpha_i a)$; from the requirement that $\alpha_1 f_1 + \cdots + \alpha_d f_d = 1$ it is immediately seen that this factor is smaller than 1.

2.2. Impact of Heterogeneity: An Approximation

As in our earlier papers [7,9], we can assess the impact of heterogeneity by parameterizing $\alpha_i = 1 + \beta_i \varepsilon$, for ε typically small; evidently, $\beta_1 f_1 + \cdots + \beta_d f_d = 0$. Relying on the Taylor series $e^x = 1 + x + x^2/2 + O(x^3)$, it is now immediate that

$$\begin{aligned} a \sum_{i=1}^{d} \alpha_i f_i e^{-\alpha_i a} &= a e^{-a} \sum_{i=1}^{d} f_i (1 + \beta_i \varepsilon) e^{-\beta_i \varepsilon a} \\ &= a e^{-a} \sum_{i=1}^{d} f_i (1 + \beta_i \varepsilon) \left(1 - \beta_i \varepsilon a + \frac{1}{2}(\beta_i \varepsilon a)^2\right) + O(\varepsilon^3) \\ &= a e^{-a} \left(1 + \frac{a}{2}(a-2) \sum_{i=1}^{d} f_i \beta_i^2 \varepsilon^2\right) + O(\varepsilon^3). \end{aligned} \tag{3}$$

The Kullback–Leibler divergence κ of the non-uniform probabilities $(1 + \beta_i \varepsilon)/N$ with respect to the uniform probabilities $1/N$ reads [7]

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

This suggests the approximation

$$\mathbb{E}S \approx k e^{-k/N} \left(1 + \frac{k}{N}\left(\frac{k}{N} - 2\right)\kappa\right). \tag{4}$$

It can even be computed what the fraction γ_j of bins covered by j balls is, when sampling aN balls. In the homogeneous case, this leads to the known result that

$$\gamma_j = \lim_{N \to \infty} \binom{aN}{j} \left(\frac{1}{N}\right)^j \left(1 - \frac{1}{N}\right)^{aN - j} = e^{-a} \frac{a^j}{j!}.$$

https://doi.org/10.1017/S0269964811000350 Published online by Cambridge University Press

we recognize the Poisson distribution. In the heterogeneous model described above,

$$\gamma_j = \sum_{i=1}^{d} f_i e^{-\alpha_i a} \frac{(\alpha_i a)^j}{j!}.$$

Note that the γ_j sum to 1, as desired (recall they represent *fractions*). Again parameterizing $\alpha_i = 1 + \beta_i \varepsilon$, for ε small, we obtain

$$\begin{aligned}\gamma_j &= \sum_{i=1}^{d} f_i e^{-a} \frac{(1 - \beta_i \varepsilon a + \frac{1}{2}(\beta_i \varepsilon a)^2)(1 + j\beta_i \varepsilon + \frac{1}{2} j(j-1)(\beta_i \varepsilon)^2)}{j!} + O(\varepsilon^3) \\ &= e^{-a} \frac{a^j}{j!} \left(1 + (a^2 + j(j-1) - 2ja) \cdot \frac{1}{2} \sum_{i=1}^{d} f_i \beta_i^2 \varepsilon^2\right) + O(\varepsilon^3).\end{aligned}$$

Replacing a by k/N and $\frac{1}{2}\sum_{i=1}^{d} f_i \beta_i^2 \varepsilon^2$ by κ, this gives an approximation for the fraction of bins covered by j balls, as a function of k, N and the 'non-homogeneity' κ:

$$\gamma_j \approx e^{-k/N} \frac{(k/N)^j}{j!} \left(1 + \left(\frac{k^2}{N^2} + j(j-1) - 2j\frac{k}{N}\right)\kappa\right). \tag{5}$$

2.3. Remarks and Example

Remark 1: The above findings also enable us to compute the fraction ϕ_j of people being in a group of size j; cf. the concept of k-anonymity in privacy [1,11]. After an elementary computation, we obtain for $j = 1, 2, \ldots$

$$\phi_j = \frac{j\gamma_j}{\sum_{\ell=1}^{\infty} \ell \gamma_\ell} = \sum_{i=1}^{d} f_i \alpha_i e^{-\alpha_i a} \frac{(\alpha_i a)^{j-1}}{(j-1)!}.$$

As before, this can be approximated by an expression in terms of k, N and κ only, under $\alpha_i = 1 + \beta_i \varepsilon$:

$$\phi_j \approx e^{-k/N} \frac{(k/N)^{j-1}}{(j-1)!} \left(1 + \left(\frac{k^2}{N^2} + j(j-1) - 2j\frac{k}{N}\right)\kappa\right). \tag{6}$$

It is a matter of elementary calculus to verify that both the approximation of γ_j (as given in Eq. (5)) and the approximation of ϕ_j (as given in Eq. (6)) add up to 1 (summing over $j = 1, 2, \ldots$), as it should.

Remark 2: We now study for which value of k the above approximation Eq. (4) is maximized. We do so by looking at the scaled version:

$$\max_{a \geq 0} a e^{-a} \left(1 + \frac{1}{2} a(a-2)\kappa\right).$$

https://doi.org/10.1017/S0269964811000350 Published online by Cambridge University Press

This yields the first-order condition

$$e^{-a}\left((1-a)-\frac{a\kappa}{2}(a-1)(a-4)\right)=0,$$

yielding the optimizer $a=1$ (which is easily seen to be a maximizer for $\kappa<\frac{2}{3}$).

We observe that Eq. (4) first increases in k, reaches a maximum N/e at $k^\star=N$, and then decreases, with limiting value 0 as $k\to\infty$. This qualitative behavior can be understood easily. For small k, there are few singletons, as there are few samples; for large k, quite likely all possible outcomes have been sampled more than once, also resulting in a low number of singletons.

For instance, in the case of birthdays, assuming that they are equally spread over the 365 days, then sampling 365 individuals maximizes the number of identifiable objects, which is (on average) 134.

Remark 3: Equation (3) confirms the claim that (for $a\leq 2$, at least) the mean number of singletons is maximized by the uniform distribution (that is, $\beta_i=0$ for all $i=1,\ldots,d$) — this is due to the absence of a linear (in ε) term in Eq. (3).

It is observed that the mean number of singletons decreases in κ for small a (that is, $a<2$), but increases for large a (that is, $a>2$). This can be intuitively understood.

- For small a, most bins will be empty or occupied by just one or two balls. Then heterogeneity leads to a smaller number of singletons, as it increases the probability that two balls end up in the same bin.
- For large a, most bins will be occupied by multiple balls. The more heterogeneity, the larger the probability that it is actually just one ball, thus leading to more singletons.

Example 1: Consider the following (somewhat stylized) example. Suppose you have data of a set of individuals, consisting of (a) zip code and (b) age. Assume that ages range from 0 to 94, and (for the moment) that all these ages are equally likely — below we indicate how to deal with heterogeneity. Now suppose that k people share the same zip code, and that we are asked to choose the number k, so as to optimize the number of uniquely identifiable individuals.

If there is no penalty imposed on the number of zip codes introduced, it is evident that it is optimal to give any individual her or his own zip code. It is more realistic to assume that there are costs, say C, for every zip code introduced. If the set of people has size M, then about $(M/k)\cdot k\exp(-k/N)$ individuals can be uniquely identified. We are therefore faced with the optimization problem

$$\max_k Me^{-k/N}-C\frac{M}{k},$$

observe that the value of M is irrelevant when determining the optimum group size $k^\star$.

https://doi.org/10.1017/S0269964811000350 Published online by Cambridge University Press

It is a matter of an elementary computation to conclude that for $C = 1$ one should have 10 individuals per zip code; for $C = 10$, we obtain 38. Adaptation to the heterogeneous case is straightforward: then

$$Me^{-k/N}\left(1 + \frac{k}{N}\left(\frac{k}{N} - 2\right)\kappa\right) - C\frac{M}{k}$$

should be maximized.

2.4. Continuous Model

The result of Proposition. 2 can be further refined. We now present its continuous counterpart. Let $\varphi(\cdot)$ be a continuous density on $[0, 1]$, and define the probability that an arbitrary ball is put in bin i by

$$\Phi_{i,N} := \int_{(i-1)/N}^{i/N} \varphi(x)\mathrm{d}x.$$

Then, in the scaled model, due to Proposition 1,

$$\begin{aligned}\lim_{N\to\infty} \frac{\mathbb{E}S(N)}{N} &= a \lim_{N\to\infty} \sum_{i=1}^{N} (1 - \Phi_{i,N})^{aN-1}\Phi_{i,N} \\ &= a \lim_{N\to\infty} \sum_{i=1}^{N} \left(1 - \frac{1}{N}\varphi\left(\frac{i}{N}\right)\right)^{aN-1} \frac{1}{N}\varphi\left(\frac{i}{N}\right).\end{aligned}$$

Now it is a matter of straightforward analysis to derive the following result.

PROPOSITION 3: *In the scaled heterogeneous model defined above, the mean number of singletons satisfies, as* $N \to \infty$,

$$\frac{\mathbb{E}S(N)}{N} \to a\int_0^1 \varphi(x)e^{-\varphi(x)a}\mathrm{d}x.$$

Example 2: Consider the density $\varphi_\gamma(x) = \gamma(x - \frac{1}{2}) + 1$, with $\gamma \in [-2, 2]$. The substitution $y := a(\gamma(x - \frac{1}{2}) + 1)$ yields

$$a\int_0^\infty \varphi_\gamma(x)e^{-\varphi_\gamma(x)a}\mathrm{d}x = \frac{1}{\gamma a}\int_{a(1-\gamma/2)}^{a(1+\gamma/2)} ye^{-y}\mathrm{d}y.$$

After some calculus, this expression can be rewritten as

$$e^{-a}\left(\frac{a+1}{\gamma a}\right)\left(e^{a\gamma/2} - e^{-a\gamma/2}\right) - \frac{e^{-a}}{2}\left(e^{a\gamma/2} + e^{-a\gamma/2}\right).$$

https://doi.org/10.1017/S0269964811000350 Published online by Cambridge University Press

For instance, for $\gamma = 2$, we thus find

$$\frac{\mathbb{E}S(N)}{N} \to \frac{1 - e^{-2a} - 2ae^{-2a}}{2a},$$

which is maximized for $a \approx 0.90$.

We could use an approximation in the spirit of Eq. (3) to approximate $\mathbb{E}S$. For this model, it takes a straightforward computation to obtain that the Kullback–Leibler distance, as a function of the 'asymmetry parameter' γ equals

$$\kappa = \frac{1}{2\gamma}\left(\left(1+\frac{\gamma}{2}\right)^2 \log\left(1+\frac{\gamma}{2}\right) - \left(1-\frac{\gamma}{2}\right)^2 \log\left(1-\frac{\gamma}{2}\right) - 1\right).$$

Observe that κ is minimal for $\gamma = 0$ (corresponding with the uniform distribution) and symmetric around 0, as could be expected.

3. VARIANCE OF THE NUMBER OF IDENTIFIABLE OBJECTS

This section considers the variance of the number of singletons. Again, after giving exact expressions and approximations, we study the impact of heterogeneity.

3.1. Explicit Expressions

As usual, we start with the homogeneous case. Let I_j be the indicator function of the event that there is exactly one ball in bin j (where $j = 1, \ldots, N$). It was observed before that

$$\mathbb{P}(I_j = 1) = k\frac{1}{N}\left(1 - \frac{1}{N}\right)^{k-1},$$

but it is easily verified that for $j_1 \neq j_2$,

$$\mathbb{P}(I_{j_1} = 1, I_{j_2} = 1) = k(k-1)\frac{1}{N^2}\left(1 - \frac{2}{N}\right)^{k-2}.$$

Observe that $S = I_1 + \cdots + I_N$. From

$$\mathbb{V}\mathrm{ar}S = \mathbb{E}S^2 - (\mathbb{E}S)^2 = \sum_{i=1}^{N} \mathbb{E}I_i + \sum_{i \neq j} \mathbb{E}I_i I_j - (\mathbb{E}S)^2,$$

we find (noting that there are $N(N-1)$ terms for which $i \neq j$)

$$\mathbb{V}\mathrm{ar}S = k\left(1 - \frac{1}{N}\right)^{k-1} + k(k-1)\frac{N-1}{N}\left(1 - \frac{2}{N}\right)^{k-2} - k^2\left(1 - \frac{1}{N}\right)^{2k-4}.$$

https://doi.org/10.1017/S0269964811000350 Published online by Cambridge University Press

Again we can consider the random variable $S(N)$, after scaling $k \equiv aN$. Directly from the previous formula, we obtain

$$\lim_{N\to\infty} \frac{\mathbb{V}\mathrm{ar}S(N)}{N} = ae^{-a} + a^2 \lim_{N\to\infty} N\left(\left(1-\frac{2}{N}\right)^{aN} - \left(1-\frac{1}{N}\right)^{2aN}\right).$$

It is clear that

$$\lim_{N\to\infty} N\left(\left(1-\frac{2}{N}\right)^{aN} - \left(1-\frac{1}{N}\right)^{2aN}\right) = f'(0),$$

with $f(x) = (1-x)^{2a/x}$. Straightforward calculus yields that $f'(0) = -ae^{-2a}$. In other words, in the homogeneous model,

$$\lim_{N\to\infty} \frac{\mathbb{V}\mathrm{ar}S(N)}{N} = ae^{-a}(1-a^2e^{-a}).$$

We now consider the heterogeneous case. Recall the standard relation

$$\mathbb{V}\mathrm{ar}S = \sum_{i=1}^{d} \mathbb{V}\mathrm{ar}S_i + \sum_{i\neq j} \mathbb{C}\mathrm{ov}(S_i, S_j).$$

Let us first compute $\mathbb{V}\mathrm{ar}S_i = \mathbb{E}S_i^2 - (\mathbb{E}S_i)^2$. Observing that we already found $\mathbb{E}S_i$ in Eq. (2), we now focus on $\mathbb{E}S_i^2$. Conditioning on the number of objects that ends up in group i (which we assumed to have F_i elements, each with probability α_i/N) yields

$$\mathbb{E}S_i^2 = \sum_{j=0}^{k} \binom{k}{j} \left(\frac{\alpha_i F_i}{N}\right)^j \left(1-\frac{\alpha_i F_i}{N}\right)^{k-j} \mathbb{E}(S_i^2 \mid N_i = j).$$

As earlier,

$$\mathbb{E}(S_i^2 \mid N_i = j) = j\left(1-\frac{1}{F_i}\right)^{j-1} + j(j-1)\frac{F_i-1}{F_i}\left(1-\frac{2}{F_i}\right)^{j-2},$$

so that

$$\mathbb{E}S_i^2 = kF_i\left(1-\frac{\alpha_i}{N}\right)^{k-1}\frac{\alpha_i}{N} + k(k-1)F_i^2\left(1-\frac{2\alpha_i}{N}\right)^{k-2}\left(\frac{\alpha_i}{N}\right)^2. \tag{7}$$

We are now left with computing $\mathbb{C}\mathrm{ov}(S_i, S_j) = \mathbb{E}S_iS_j - \mathbb{E}S_i\,\mathbb{E}S_j$ for $i \neq j$. As we already know $\mathbb{E}S_i$, we focus on $\mathbb{E}S_iS_j$. It holds that

$$\mathbb{E}S_iS_j = \sum_{\ell_i=0}^{k}\sum_{\ell_j=0}^{k-\ell_i} \binom{k}{\ell_i,\ell_j} \left(\frac{\alpha_i F_i}{N}\right)^{\ell_i} \left(\frac{\alpha_j F_j}{N}\right)^{\ell_j} \times \left(1-\frac{\alpha_i F_i}{N}-\frac{\alpha_j F_j}{N}\right)^{k-\ell_i-\ell_j} \mathbb{E}(S_iS_j \mid N_i = \ell_i, N_j = \ell_j),$$

https://doi.org/10.1017/S0269964811000350 Published online by Cambridge University Press

and in addition a conditional independence argument yields that

$$\mathbb{E}(S_i S_j \,|\, N_i = \ell_i, N_j = \ell_j) = \ell_i \left(1 - \frac{1}{F_i}\right)^{\ell_i - 1} \ell_j \left(1 - \frac{1}{F_j}\right)^{\ell_j - 1}.$$

Standard computations now yield that

$$\mathbb{E}S_i S_j = k(k-1) F_i F_j \left(1 - \frac{\alpha_i}{N} - \frac{\alpha_j}{N}\right)^{k-2} \frac{\alpha_i}{N} \frac{\alpha_j}{N}. \tag{8}$$

Now all the above findings can be collected.

PROPOSITION 4: *In the heterogeneous model defined above, the variance of the number of singletons equals*

$$\mathbb{V}\mathrm{ar}S = \sum_{i=1}^{d} \left(\mathbb{E}S_i^2 - (\mathbb{E}S_i)^2\right) + \sum_{i \neq j} \left(\mathbb{E}S_i S_j - \mathbb{E}S_i \, \mathbb{E}S_j\right),$$

with $\mathbb{E}S_i$ *given by* (2), $\mathbb{E}S_i^2$ *by* (7) *and* $\mathbb{E}S_i S_j$ *by* (8).

We now again look at the scaled variant. As before,

$$\frac{\mathbb{V}\mathrm{ar}S_i(N)}{N} \to a\alpha_i f_i e^{-\alpha_i a} \left(1 - a^2 \alpha_i f_i e^{-\alpha_i a}\right).$$

Also, due to Lemma 1,

$$\frac{\mathbb{C}\mathrm{ov}(S_i(N), S_j(N))}{N} \to -a^3 \alpha_i^2 f_i \alpha_j^2 f_j e^{-(\alpha_i + \alpha_j)a}.$$

We arrive at the following statement.

PROPOSITION 5: *In the scaled heterogeneous model defined above, the variance of the number of singletons satisfies, as* $N \to \infty$,

$$\begin{aligned}
\frac{\mathbb{V}\mathrm{ar}S(N)}{N} &\to \sum_{i=1}^{d} a\alpha_i f_i e^{-\alpha_i a} \left(1 - a^2 \alpha_i f_i e^{-\alpha_i a}\right) - \sum_{i \neq j} a^3 \alpha_i^2 f_i \alpha_j^2 f_j e^{-(\alpha_i + \alpha_j)a} \\
&= a \sum_{i=1}^{d} \alpha_i f_i e^{-\alpha_i a} - a^3 \sum_{i=1}^{d} \sum_{j=1}^{d} \alpha_i^2 f_i \alpha_j^2 f_j e^{-(\alpha_i + \alpha_j)a} \\
&= a \sum_{i=1}^{d} \alpha_i f_i e^{-\alpha_i a} - a^3 \left(\sum_{i=1}^{d} \alpha_i^2 f_i e^{-\alpha_i a}\right)^2.
\end{aligned}$$

https://doi.org/10.1017/S0269964811000350 Published online by Cambridge University Press

3.2. Impact of Heterogeneity: An Approximation

We again parameterize $\alpha_i = 1 + \beta_i \varepsilon$. We already observed that

$$a \sum_{i=1}^{d} \alpha_i f_i e^{-\alpha_i a} = a e^{-a} \left(1 + \frac{a}{2}(a-2) \sum_{i=1}^{d} f_i \beta_i^2 \varepsilon^2 \right) + O(\varepsilon^3),$$

whereas it turns out that

$$a^3 \left(\sum_{i=1}^{d} \alpha_i^2 f_i e^{-\alpha_i a} \right)^2 = a^3 e^{-2a} \left(1 + (2-3a) \sum_{i=1}^{d} f_i \beta_i^2 \varepsilon^2 \right) + O(\varepsilon^3).$$

This leads to the approximation (for the unscaled model)

$$\mathbb{V}\mathrm{ar}S \approx k e^{-k/N} \left(1 - \frac{k}{N} \left(\frac{k}{N} - 2 \right) \kappa \right) - \frac{k^3}{N^2} e^{-2k/N} \left(1 + \left(4 - 6\frac{K}{N} \right) \kappa \right).$$

3.3. Continuous Model

We now consider the continuous model, analogously to Section 2.4; the probability of a ball being put in bin i is $\Phi_{i,N}$, equalling the integral over the density $\varphi(\cdot)$ between $(i-1)/N$ and i/N, for $i = 1, \ldots, N$. The proof of following result is similar to the proof of Proposition 3.

PROPOSITION 6: *In the scaled heterogeneous model defined above, the variance of the number of singletons satisfies, as* $N \to \infty$,

$$\frac{\mathbb{V}\mathrm{ar}S(N)}{N} \to \int_0^1 a\varphi(x)(1 - a^2\varphi(x)) e^{-\varphi(x)a} \mathrm{d}x.$$

We conjecture that $(S(N) - \mathbb{E}S(N))/\sqrt{\mathbb{V}\mathrm{ar}S(N)/N}$ converges to a standard Normal random variable.

4. PROBABILITY OF AT LEAST ONE IDENTIFIABLE OBJECT

Let $\xi(k, N)$ be the probability of at least one identifiable object, that is, the probability $\mathbb{P}(S > 0)$ of at least one singleton. Particularly if k is large relative to N, this is an interesting anonymity metric. (An example could be: suppose you receive data about the ages of 300 people; is there anyone among these 300 people whose age is unique?). In this section we develop a recursive scheme to evaluate $\xi(k, N)$.

https://doi.org/10.1017/S0269964811000350 Published online by Cambridge University Press

4.1. Recursive Scheme

We analyze this probability by computing the probability $\zeta(k,N)$ of its complement (that is, *no* singletons); we start with the homogeneous case. Consider an arbitrary ball that ends up in an arbitrary bin. As there should not be singletons, it means that at least one more ball (out of the remaining $k-1$) should be in that bin as well; realize that the number of balls that are in that bin (apart from the tagged one) follows a binomial distribution with parameters $k-1$ and $1/N$. We thus find

$$\zeta(k,N) = \sum_{j=1}^{k-1} \binom{k-1}{j} \left(\frac{1}{N}\right)^j \left(1-\frac{1}{N}\right)^{k-1-j} \zeta(k-1-j,N-1).$$

The initialization of this recursion is $\zeta(k,1)=1$ and $1-\zeta(0,N)=\zeta(1,N)=0$ for any $k=2,3,ldots$ and $N=1,2,\ldots$. The first steps can be performed easily:

$$\zeta(2,N) = \frac{1}{N}, \quad \zeta(3,N) = \frac{1}{N^2}, \quad \zeta(4,N) = \frac{3N-2}{N^3},$$

and, with a bit more effort,

$$\zeta(5,N) = \frac{10N-9}{N^4}, \quad \zeta(6,N) = \frac{15N^2-20N+6}{N^5}$$
$$\zeta(7,N) = \frac{105N^2-259N+155}{N^6}$$

Table 1 presents the values of $\zeta(k,N)$ for $k=1,\ldots,50$ and $N=1,\ldots,20$. It shows that $\zeta(k,N)$ goes to 1 for k large. In addition, for fixed k, $\zeta(k,N)$ decreases with N. A nice sanity check for formulae for $\zeta(k,N)$ is the relation ($k \geq 3$)

$$\zeta(k,2) = 1 - \frac{k}{2^{k-1}}.$$

4.2. Full Distribution of Number of Singletons

The above results immediately lead to the full distribution of the number of singletons S; for ease, we restrict ourselves to the uniform case. It is seen that

$$\mathbb{P}(S=j) = \binom{N}{j} k(k-1)\cdots(k-j+1) \left(\frac{1}{N}\right)^j \left(1-\frac{j}{N}\right)^{k-j} \zeta(k-j,N-j),$$

evidently $\mathbb{P}(S=0)=\zeta(k,N)$. Evidently, for $k \leq N$, we already knew from the standard birthday problem that

$$\mathbb{P}(S=k) = \frac{N!/(N-k)!}{N^k}.$$

https://doi.org/10.1017/S0269964811000350 Published online by Cambridge University Press

https://doi.org/10.1017/S0269964811000350 Published online by Cambridge University Press

TABLE 1. $\zeta(k,N)$ for $k = 1,\ldots,50$ and $N = 1,\ldots,20$

k/N	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
2	1.000	0.500	0.333	0.250	0.200	0.167	0.143	0.125	0.111	0.100	0.091	0.083	0.077	0.071	0.067	0.063	0.059	0.056	0.053	0.050
3	1.000	0.250	0.111	0.063	0.040	0.028	0.020	0.016	0.012	0.010	0.008	0.007	0.006	0.005	0.004	0.004	0.003	0.003	0.003	0.003
4	1.000	0.500	0.259	0.156	0.104	0.074	0.055	0.043	0.034	0.028	0.023	0.020	0.017	0.015	0.013	0.011	0.010	0.009	0.008	0.007
5	1.000	0.688	0.259	0.121	0.066	0.039	0.025	0.017	0.012	0.009	0.007	0.005	0.004	0.003	0.003	0.002	0.002	0.002	0.001	0.001
6	1.000	0.813	0.333	0.162	0.090	0.055	0.036	0.025	0.018	0.013	0.010	0.008	0.006	0.005	0.004	0.003	0.003	0.002	0.002	0.002
7	1.000	0.891	0.443	0.195	0.095	0.051	0.030	0.018	0.012	0.008	0.006	0.004	0.003	0.002	0.002	0.001	0.001	0.001	0.001	0.001
8	1.000	0.938	0.557	0.240	0.114	0.060	0.034	0.021	0.013	0.009	0.006	0.004	0.003	0.002	0.002	0.001	0.001	0.001	0.001	0.001
9	1.000	0.965	0.660	0.302	0.139	0.069	0.036	0.021	0.012	0.008	0.005	0.003	0.002	0.002	0.001	0.001	0.001	0.001	0.000	0.000
10	1.000	0.980	0.744	0.378	0.171	0.081	0.042	0.023	0.013	0.008	0.005	0.003	0.002	0.002	0.001	0.001	0.001	0.000	0.000	0.000
11	1.000	0.989	0.811	0.460	0.211	0.099	0.049	0.026	0.014	0.008	0.005	0.003	0.002	0.001	0.001	0.001	0.000	0.000	0.000	0.000
12	1.000	0.994	0.862	0.541	0.261	0.121	0.058	0.029	0.016	0.009	0.005	0.003	0.002	0.001	0.001	0.001	0.000	0.000	0.000	0.000
13	1.000	0.997	0.900	0.617	0.319	0.148	0.070	0.035	0.018	0.010	0.006	0.003	0.002	0.001	0.001	0.001	0.000	0.000	0.000	0.000
14	1.000	0.998	0.928	0.684	0.382	0.182	0.085	0.041	0.021	0.011	0.006	0.004	0.002	0.001	0.001	0.001	0.000	0.000	0.000	0.000
15	1.000	0.999	0.949	0.742	0.446	0.222	0.104	0.050	0.025	0.013	0.007	0.004	0.002	0.001	0.001	0.001	0.000	0.000	0.000	0.000
16	1.000	1.000	0.963	0.792	0.510	0.267	0.127	0.060	0.029	0.015	0.008	0.004	0.002	0.001	0.001	0.001	0.000	0.000	0.000	0.000
17	1.000	1.000	0.974	0.833	0.572	0.317	0.155	0.073	0.035	0.017	0.009	0.005	0.003	0.001	0.001	0.001	0.000	0.000	0.000	0.000
18	1.000	1.000	0.982	0.866	0.629	0.369	0.187	0.089	0.043	0.021	0.010	0.005	0.003	0.002	0.001	0.001	0.000	0.000	0.000	0.000
19	1.000	1.000	0.987	0.894	0.681	0.423	0.223	0.108	0.052	0.025	0.012	0.006	0.003	0.002	0.001	0.001	0.000	0.000	0.000	0.000
20	1.000	1.000	0.991	0.916	0.727	0.476	0.263	0.131	0.063	0.030	0.015	0.007	0.004	0.002	0.001	0.001	0.000	0.000	0.000	0.000
21	1.000	1.000	0.994	0.934	0.768	0.528	0.306	0.157	0.076	0.036	0.018	0.009	0.004	0.002	0.001	0.001	0.000	0.000	0.000	0.000
22	1.000	1.000	0.996	0.948	0.804	0.578	0.350	0.186	0.092	0.044	0.021	0.010	0.005	0.003	0.001	0.001	0.000	0.000	0.000	0.000
23	1.000	1.000	0.997	0.959	0.835	0.625	0.396	0.218	0.110	0.053	0.026	0.013	0.006	0.003	0.002	0.001	0.000	0.000	0.000	0.000
24	1.000	1.000	0.998	0.968	0.861	0.668	0.442	0.253	0.131	0.064	0.031	0.015	0.007	0.004	0.002	0.001	0.001	0.000	0.000	0.000
25	1.000	1.000	0.999	0.975	0.884	0.707	0.487	0.291	0.155	0.077	0.038	0.018	0.009	0.004	0.002	0.001	0.001	0.000	0.000	0.000
26	1.000	1.000	0.999	0.980	0.903	0.743	0.531	0.329	0.181	0.092	0.045	0.022	0.011	0.005	0.003	0.001	0.001	0.000	0.000	0.000
27	1.000	1.000	0.999	0.985	0.919	0.776	0.574	0.369	0.210	0.110	0.054	0.026	0.013	0.006	0.003	0.002	0.001	0.000	0.000	0.000
28	1.000	1.000	1.000	0.988	0.933	0.804	0.614	0.409	0.241	0.129	0.065	0.032	0.015	0.008	0.004	0.002	0.001	0.000	0.000	0.000
29	1.000	1.000	1.000	0.991	0.944	0.830	0.651	0.449	0.274	0.151	0.077	0.038	0.019	0.009	0.004	0.002	0.001	0.001	0.000	0.000
30	1.000	1.000	1.000	0.993	0.954	0.853	0.686	0.489	0.308	0.174	0.092	0.046	0.022	0.011	0.005	0.003	0.001	0.001	0.000	0.000

(*continued*)

https://doi.org/10.1017/S0269964811000350 Published online by Cambridge University Press

Table 1. Continued

k/N	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
31	1.000	1.000	1.000	0.994	0.962	0.872	0.719	0.527	0.343	0.200	0.107	0.055	0.027	0.013	0.006	0.003	0.002	0.001	0.000	0.000
32	1.000	1.000	1.000	0.996	0.968	0.890	0.748	0.564	0.378	0.227	0.125	0.065	0.032	0.016	0.008	0.004	0.002	0.001	0.000	0.000
33	1.000	1.000	1.000	0.997	0.974	0.905	0.775	0.599	0.414	0.256	0.145	0.076	0.039	0.019	0.009	0.005	0.002	0.001	0.001	0.000
34	1.000	1.000	1.000	0.997	0.978	0.918	0.800	0.633	0.449	0.286	0.166	0.089	0.046	0.023	0.011	0.005	0.003	0.001	0.001	0.000
35	1.000	1.000	1.000	0.998	0.982	0.930	0.822	0.664	0.484	0.317	0.189	0.104	0.054	0.027	0.013	0.007	0.003	0.002	0.001	0.000
36	1.000	1.000	1.000	0.998	0.985	0.940	0.842	0.694	0.518	0.349	0.213	0.120	0.064	0.032	0.016	0.008	0.004	0.002	0.001	0.000
37	1.000	1.000	1.000	0.999	0.988	0.948	0.860	0.722	0.551	0.381	0.239	0.138	0.074	0.038	0.019	0.009	0.005	0.002	0.001	0.001
38	1.000	1.000	1.000	0.999	0.990	0.956	0.877	0.747	0.583	0.413	0.266	0.157	0.086	0.045	0.023	0.011	0.006	0.003	0.001	0.001
39	1.000	1.000	1.000	0.999	0.992	0.962	0.891	0.771	0.614	0.445	0.293	0.178	0.100	0.053	0.027	0.013	0.007	0.003	0.002	0.001
40	1.000	1.000	1.000	0.999	0.993	0.967	0.904	0.793	0.643	0.476	0.322	0.199	0.114	0.062	0.032	0.016	0.008	0.004	0.002	0.001
41	1.000	1.000	1.000	1.000	0.995	0.972	0.915	0.813	0.670	0.507	0.351	0.222	0.130	0.072	0.038	0.019	0.010	0.005	0.002	0.001
42	1.000	1.000	1.000	1.000	0.996	0.976	0.925	0.831	0.696	0.537	0.380	0.246	0.148	0.083	0.044	0.023	0.011	0.006	0.003	0.001
43	1.000	1.000	1.000	1.000	0.996	0.980	0.934	0.848	0.721	0.566	0.409	0.271	0.166	0.095	0.051	0.027	0.013	0.007	0.003	0.002
44	1.000	1.000	1.000	1.000	0.997	0.983	0.942	0.863	0.743	0.594	0.438	0.297	0.186	0.108	0.060	0.031	0.016	0.008	0.004	0.002
45	1.000	1.000	1.000	1.000	0.998	0.985	0.949	0.877	0.765	0.621	0.466	0.323	0.206	0.123	0.069	0.037	0.019	0.010	0.005	0.002
46	1.000	1.000	1.000	1.000	0.998	0.987	0.956	0.890	0.785	0.647	0.494	0.349	0.228	0.138	0.079	0.043	0.022	0.011	0.006	0.003
47	1.000	1.000	1.000	1.000	0.998	0.989	0.961	0.901	0.803	0.671	0.522	0.376	0.250	0.155	0.090	0.050	0.026	0.013	0.007	0.003
48	1.000	1.000	1.000	1.000	0.999	0.991	0.966	0.911	0.820	0.695	0.549	0.402	0.273	0.173	0.102	0.057	0.031	0.016	0.008	0.004
49	1.000	1.000	1.000	1.000	0.999	0.992	0.970	0.921	0.836	0.717	0.575	0.429	0.297	0.191	0.115	0.066	0.036	0.019	0.009	0.005
50	1.000	1.000	1.000	1.000	0.999	0.993	0.974	0.929	0.850	0.738	0.600	0.455	0.321	0.211	0.129	0.075	0.041	0.022	0.011	0.006

4.3. Heterogeneous case

Once we have computed the numbers $\zeta(i,j)$ (that correspond to the homogeneous case), it is fairly easy to deal with the heterogeneous case:

$$\zeta(k,N) = \sum_{j} \binom{k}{j_1,\ldots,j_d} \prod_{i=1}^{d} \left(\frac{\alpha_i F_i}{N}\right)^{j_i} \zeta_{\mathrm{u}}(j_i, F_i),$$

where $\zeta_{\mathrm{u}}(\cdot\,,\cdot)$ refers to the probability of no singletons in the uniform case, and the summation is over vectors $\boldsymbol{j} \in \{0,1,\ldots\}^d$ such that $j_1 + \cdots + j_d = k$. It is observed that this expression is hard to evaluate, as one has to sum over all vectors $\boldsymbol{j}$ whose entries add up to k, whose number grows explosively in k. This explains the need for approximations. One such approximation relies on the idea of replacing the multinomial distribution by the corresponding Poisson distribution (where the individual components are assumed to be independent). Then one obtains

$$\zeta(k,N) = \prod_{i=1}^{d} \left(\sum_{j=0}^{\infty} \left(\exp\left(-\frac{k\alpha_i F_i}{N}\right) \left(\frac{k\alpha_i F_i}{N}\right)^{j} \Big/ j! \right) \zeta_{\mathrm{u}}(j, F_i) \right).$$

5. NUMERICAL EXPERIMENTS

In this section, we report on numerical experiments that we performed with demographic data of all 428 Dutch municipalities. For all of them we first determined the distribution of age over the population (truncating at 94 years, leaving us with 95 bins, viz. 0 up to and including 94). Then, we computed on the basis of these data the Kullback–Leibler distance κ, and the mean and variance of the number of singletons. The mean $\mathbb{E}S$ and the mean plus/minus twice the standard error $\mathbb{E}S \pm 2\sqrt{\mathbb{V}\mathrm{ar}S}$ are depicted in the left panels of Figure 1, as a function of the κ — each dot represents one municipality.

We also include approximation Eq. (4), which is a linear function of κ. As argued in the derivation, it is supposed to perform well if the distance with respect to the uniform distribution is relatively modest. From the left panels of Figure 2, it is seen that the approximation does not give an accurate prediction. This is mainly because the distribution is highly non-uniform for the higher ages (ages above, say, 85 are hardly represented). In the right panels, we depicted the same experiments, but just for the ages 0 up to and including 79, and there we indeed see an excellent fit.

Although the left panels indicate that approximation Eq. (4) does not yield an accurate estimate for the mean number of singletons $\mathbb{E}S$ in-case the non-uniformity is too strong, the (nearly) linear shape of the scatter plot does show that knowledge of the Kullback–Leibler distance accurately predicts $\mathbb{E}S$. One could for instance approximate $\mathbb{E}S$ (as a function of κ) by the linear regression $\delta_0 + \delta_1\kappa$, where δ_0 and δ_1 are estimated by a least squares procedure.

https://doi.org/10.1017/S0269964811000350 Published online by Cambridge University Press

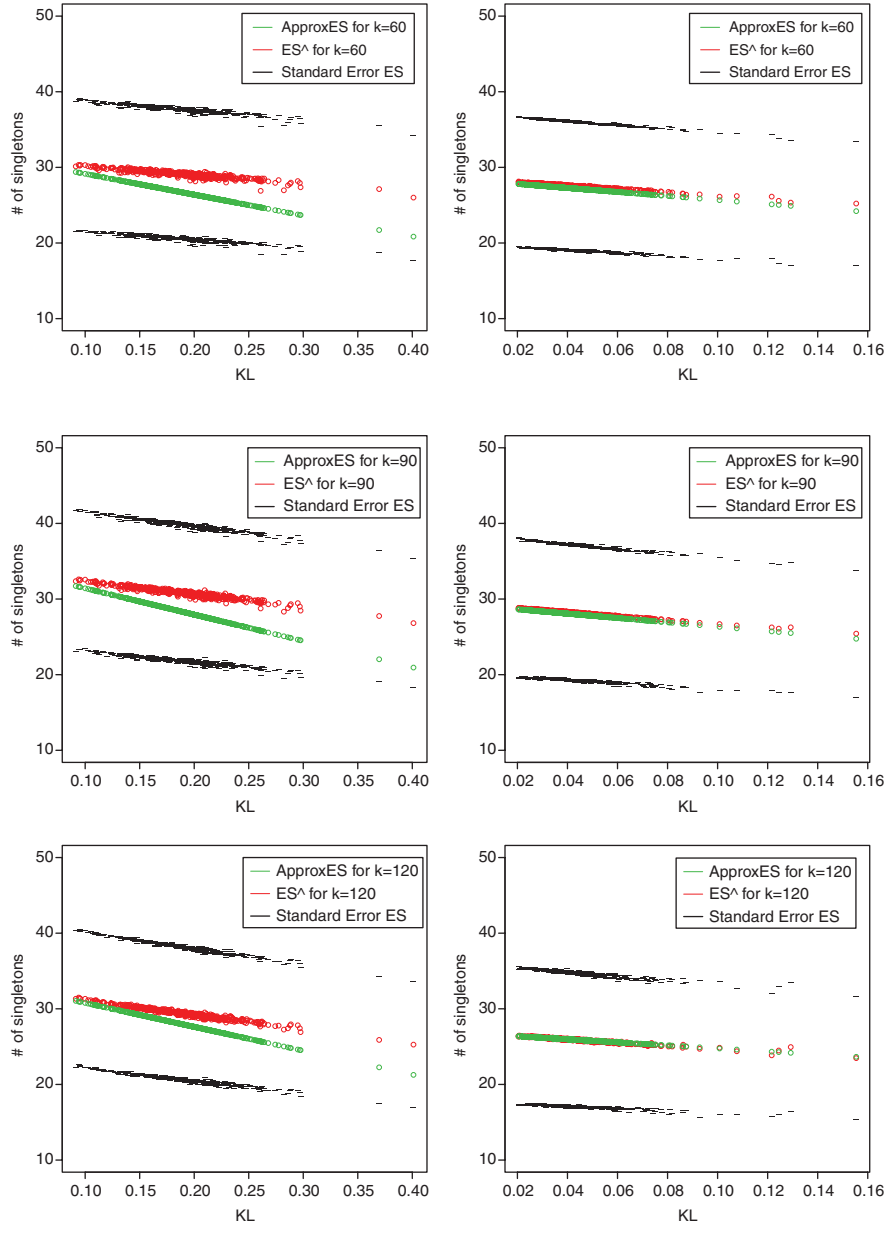


FIGURE 1. (Color online) Mean number of singletons, as a function of the Kullback–Leibler distance κ. Left panels: full population; right panels: ages 0–79 only. Top to bottom: $k = 60, 90$ and 120.

https://doi.org/10.1017/S0269964811000350 Published online by Cambridge University Press

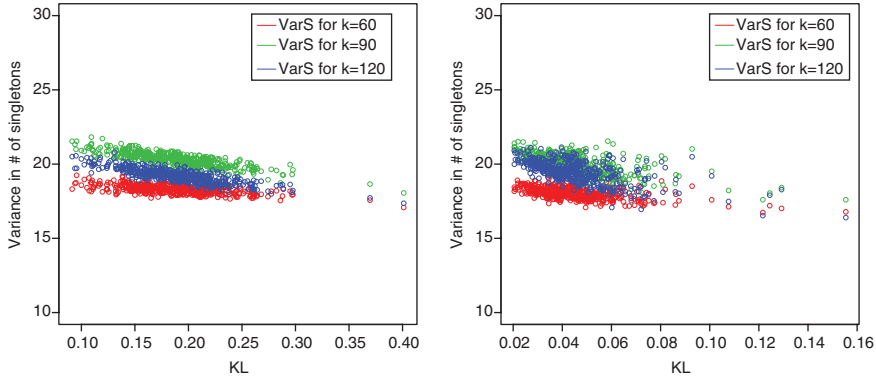


FIGURE 2. (Color online) Variance of the number of singletons, as a function of the Kullback–Leibler distance κ. Left panel: full population; right panel: ages 0–79 only.

The left panels show that the mean number of singletons is highest for $k = 90$, which could be expected from Remark 2 (recall that $N = 95$ here). Additional experiments (not reported on here) show that when leaving out the ages 80–94, the mean number of singletons is indeed highest around $k = 80$.

Figure 2 shows a scatter plot of the Kullback–Leibler distance κ and the variance $\mathbb{V}\mathrm{ar}S$. For the full population, we observe three decreasing, more or less linear lines; when leaving out the ages 79–94 there is hardly any sensitivity in κ.

6. CONCLUDING REMARKS

This paper presented an analysis of the number of singletons in the setting of the generalized birthday problem. Various metrics have been studied. Special attention has been paid to obtaining insight into the impact of heterogeneity on the number of singletons. Future research includes extensive testing with demographic data.

Acknowledgments

Sasha Gnedin (University of Utrecht) is thanked for useful discussions. The research of M. Koot was in part supported by the FES 2009 ICT project COMMIT and the VL-e project.

References

1. Aggarwal, C. (2005). On k-anonymity and the curse of dimensionality. *Proceedings of VLDB 2005*, pp. 901–909.
2. Dembo, A. & Zeitouni, O. (1998). *Large deviations techniques and applications*, 2nd ed. New York: Springer.

https://doi.org/10.1017/S0269964811000350 Published online by Cambridge University Press

3. Diaconis, P. & Mosteller, F. (1989). Methods for studying coincidences. *Journal of the American Statistical Association* 84: 853–861.
4. 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.
5. Joag-Dev, K. & Proschan, F. (1992). The birthday problem with unlike probabilities. *American Mathematical Monthly* 99: 10–12.
6. Klotz, J. (1979). The birthday problem with unequal probabilities. Technical Report, No. 59. Department of Statistics, University of Wisconsin.
7. Koot, M., Mandjes, M., & De Laat, C. (2011). Efficient probabilistic estimation of quasi-identifier uniqueness. Proceedings of NWO ICT. Open 2011.
8. Kullback, S. & Leibler, R. (1951). On information and sufficiency. *Annals of Mathematical Statistics* 22: 79–86.
9. Mandjes, M. (2011). Generalized birthday problems in the large-deviations regime. *Submitted.*
10. Nunnikhoven, T. (1992). A birthday problem solution for non-uniform birth frequencies. *American Statistician* 46: 270–274.
11. Sweeney, L. (2002). *k*-anonymity: a model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems* 10: 557–570.

APPENDIX: A Useful Lemma

LEMMA 1: *In the scaled heterogeneous model, as* $N \to \infty$,

$$\frac{\mathbb{C}\mathrm{ov}(S_i(N), S_j(N))}{N} \to -a^3\alpha_i^2 f_i \alpha_j^2 f_j e^{-(\alpha_i+\alpha_j)a}.$$

PROOF: From the expressions in Section 3, it is straightforward that

$$\begin{aligned}
&\frac{\mathbb{C}\mathrm{ov}(S_i(N), S_j(N))}{N(\alpha_i f_i\, \alpha_j f_j)} \\
&= a(Na-1)\left(1-\frac{\alpha_i+\alpha_j}{N}\right)^{Na} - a^2 N\left(1-\frac{\alpha_i}{N}\right)^{Na}\left(1-\frac{\alpha_j}{N}\right)^{Na} \\
&\sim a^2 N\left(\left(1-\frac{\alpha_i+\alpha_j}{N}\right)^{Na} - \left(1-\frac{\alpha_i}{N}\right)^{Na}\left(1-\frac{\alpha_j}{N}\right)^{Na}\right),
\end{aligned}$$

where $f(n) \sim g(n)$ denotes that $f(n)/g(n) \to 1$ as $n \to \infty$. We have, due to L'Hôptital's rule, for $A, B \in \mathbb{R}$,

$$\lim_{N\to\infty} \frac{\left(1-\frac{A+B}{N}\right)^{Na} - \left(1-\frac{A}{N}\right)^{Na}\left(1-\frac{B}{N}\right)^{Na}}{\frac{1}{N}} = \psi'(0),$$

with $\psi(x) := (1-(A+B)x)^{a/x} - (1-Ax)^{a/x}(1-Bx)^{a/x}$. Using Taylor expansions, we find $\psi'(0) = -aABe^{-a(A+B)}$. Now plugging in $A = \alpha_i$ and $B = \alpha_j$ yields the stated result. ■

https://doi.org/10.1017/S0269964811000350 Published online by Cambridge University Press