Flaws in the UIDAI Process

HANS VERGHESE MATHEWS

The accuracy of biometric identification depends on the chance of a false positive: the probability that the identifiers of two persons will match. Individuals whose identifiers match might be termed duplicands. When very many people are to be identified success can be measured by the (low) proportion of duplicands. The Government of India is engaged upon biometrically identifying the entire population of India. An experiment performed at an early stage of the programme has allowed us to estimate the chance of a false positive: and from that to estimate the proportion of duplicands. For the current population of 1.2 billion the expected proportion of duplicands is 1/121, a ratio which is far too high.

A legal challenge is being mounted in the Supreme Court, currently, to the programme of biometric identification that the Unique Identification Authority of India (UIDAI) is engaged upon: an identification preliminary and requisite to providing citizens with "Aadhaar numbers" that can serve them as "unique identifiers" in their transactions with the state. What follows will recount an assessment of their chances of success. We shall be using data that was available to the UIDAI and shall employ only elementary ways of calculation. It should be recorded immediately that an earlier technical paper by the author (Mathews 2013) has been of some use to the plaintiffs; and reference will be made to that in due course.

The Aadhaar numbers themselves may or may not derive, in some way, from the biometrics in question; the question is not material here. For our purposes a *biometric* is a *numerical representation* of some organic feature: like the iris or the retina, for instance, or the inside of a finger, or the hand taken whole even. We shall consider them in some more detail later. The UIDAI is using fingerprints and iris images to generate a combination of biometrics for each individual. This paper bears on the accuracy of the composite biometric identifier. How well those composites will distinguish between individuals can be assessed, actually, using the results of an experiment conducted by the UIDAI itself in the very early stages of its operation; and our contention is that, from those results themselves, the UIDAI should have been able to estimate *how many individuals would have their biometric identifiers matching those of some other person*, under the best of circumstances even, when any good part of population has been identified. Let us term a *duplicand* any person whose biometric identifier matches that of some other person: an occurrence which can by no means be ruled out. In the best of circumstances no citizen would try to obtain more than one Aadhaar number; and Table 1 lists the number of duplicands to be expected even in such circumstances, and conservatively expected when the population will almost all have been identified.

The last column of Table 1 lists the ratio of the second column to the first; and the third row of the table shows that, given the current population of India, the UIDAI should expect one in every 121 persons to be a duplicand. The last row discloses that, by the time the population reaches 1.5 billion, they should expect

Table 1: Duplicands Expected

Identified (in billions) (1)	Duplicands Expected (2)	Duplicands/ Identified (3)
1.0	68,87,324	1/145
1.1	83,35,451	1/132
1.2	99,21,995	1/121
1.3	1,16,47,035	1/112
1.4	1,35,10,651	1/104
1.5	1,55,12,927	1/97

The author thanks Nico Temme of the Centrum Wiskunde & Informatica in The Netherlands for the bounds he derived on the chance of a false positive. He is particularly grateful to the anonymous referee of this journal who, through two rounds of comment, has very much improved the presentation of the results.
A technical supplement to this paper is placed on the EPW website along with this paper.

Hans Verghese Mathews (*hans@cis-india.org*) does mathematical and statistical modelling for the Centre for Internet and Society, Bengaluru, where he is a Fellow.

one person in every 97 to be a duplicand; and biometric identifiers which allow such high proportions of duplicands cannot be supposed, at all, to uniquely identify individuals. Given the result of the experiment conducted by the UIDAI, which was reported in [2] and which we shall presently describe, the calculation of the numbers in the second column is not difficult: requires no more calculus than anyone who has studied science or engineering would know.

A biometric is a numerical representation of an organic feature we said, which we shall regard as the *output* of some device upon it being presented as *input,* an instance of the feature being represented. Let us take the iris as our feature now. The design of the device will be attuned to the particularities of the iris considered generically. But we shall regard the device as a "black box," simply, whose interior we are not concerned with: except to note that *the output is almost never identical when any particular iris is presented as input on different occasions*. The reasons for such inconvenient disparity are not germane here. But the important consequence is that one must decide how similar two outputs must be to count as representations of one and the same input iris. The usual solution is to so construct the numerical representations that a "distance" can be calculated between any pair of them—calculated in one way for all possible pairs of course—and to take a pair of such to "match" if the distance between them does not exceed a specified "threshold." This matching threshold can be decided by experiment only; and there is always a chance now that the numerical representations of different irises may match.

The complication will extend to the composite biometric identifiers we are considering. One may ask what the chance of a match is when the individuals are known to be different. Let us call individuals the "organic sources" of our biometric identifiers; and abbreviate as *SX* the organic source of an identifier *X*. The abbreviation will prove useful as we proceed. To save writing let us term the calculation of the distance between identifiers a "comparison." Let $X \sim Y$ abbreviate a match of identifiers upon their comparison. The occurrence of a match when the organic sources are different is usually termed a false positive; and the probability of a false positive determines the accuracy of a biometric identifier. It is usual to regard this conditional probability as invariant: to take for one and the same real number, lying between 0 and 1 of course, the probability of a match between identifiers deriving from *any* pair of different organic sources. The features commonly biometrised are presumably such as to warrant the assumption.

Let ξ be the probability of a false positive for the composite biometric identifier that the UIDAI is using: of which we need an estimate to perform the calculations that give us our table. Let us consider the situation when the UIDAI has biometrcially identified some *k* different individuals: whose identifiers $Y_1, Y_2, \ldots, Y_k$ are stored in some database D say. It is usual to call these stored identifers "templates," and to say that the organic sources identified thus have been "enrolled;" so in Table 1, the first column lists the enrolled population. When the next enrollee comes along there is some chance now that his or her identifier will match *some or other* stored template: which will increase with the number already enrolled, of course, since a new identifier must be compared to each stored template. So write $\Phi(k)$ for this probability. Let *X* be the biometric identifier of the new enrollee: who is its organic source *SX* now of course. The sources $SX, SY_1, SY_2, \ldots, SY_k$ are distinct persons now: and ξ is uniformly the chance of a match $X \sim Y_i$ between *X* and any stored template Y_i in D, and $(1 - \xi)$ the probability that the match does not occur. It is usual to assume that the occurrence or not of $X \sim Y_i$ is *independent* of the like for $X \sim Y_j$ when Y_i and Y_j are different stored templates, whether or not they themselves match; and we shall suppose so. The chance that *X* will match none of the *k* stored templates is now the product $(1 - \xi)^k$ of the *k* identical probabilities that no match will occur; and as *X* will *either* match some or other among the stored templates, or match none at all, we have

$$1 - \Phi(k) = (1 - \xi)^k \qquad \ldots(1)$$

The probability $\Phi(k)$ can be reliably estimated when *k* is large enough; and solving for ξ above would yield us an estimate for the chance of a false positive. An exact such solution of (1) is readily specified via $\xi = 1 - [1 - \Phi(k)]^{1/k}$ of course; but taking *k*-th roots for large *k* is computationally intractable. Computable bounds for ξ in terms of $\Phi(k)$ itself are easily obtained though; and we have

$$\Phi(k) / k \leq \xi \leq - \log[1 - \Phi(k)] / k \qquad \ldots(2)$$

To derive the bounds (2) from the relation (1) is an elementary exercise: requiring no more than those parts of the differential calculus that students of science or engineering will have mastered in their first year, if not before.

Let *n* be the total number of individuals who are to be enrolled. We must have $n \geq 2$ here, of course, and each template will have been compared to every other when the enrolment is complete: so $n(n - 1)/2$ comparisons will have been made, altogether, between the constitutents of those many distinct pairs of biometric identifiers. Let $M(n)$ be the number of matches one must expect upon these many comparisons. Since ξ is the uniform probability of a match upon any one comparison, we have

$$M(n) = \xi\, [n(n - 1)/2] \qquad \ldots(3)$$

Now for the expected number of matches when a total of *n* different individuals have been enrolled. Table 1 does not list the expected number of matches, for varying totals, but the number of duplicands rather: enrolled persons whose identifiers are expected to match those of some or other enrolled person. Let $W(n)$ be the number of duplicands when *n* different persons have been enrolled. Table 1 lists in its second column the expected numbers $W(n)$ for the different values of *n* listed in the first column. Suppose it happens that no identifier matches *more than one* other identifier when the comparisons are all done: then each match will involve a distinct pair of identifiers and we get $W(n) = 2M(n)$. But things may not fall out so nicely. We may have matches $X \sim Y$ and $Y \sim Z$ and $Z \sim X$ between the identifiers of three different organic sources *SX*, *SY* and *SZ* for instance: in which case these three matches involve three persons, only, not six. We can limit the miscounting we might do by doubling the number of matches given by the formula (3); and we shall presently see how to do so more closely. But in the case at hand at any rate, with ξ as it happens to be for the composite biometric identifier being used by the UIDAI, it was demonstrated in [1] that

(†) *only rarely* would the identifier of a duplicand match *more than one* other identifier.

The computations in [1] had incorporated the probability that indviduals would try to enrol more than once: which had been estimated as 0.0005 by the UIDAI in their report. The result (†) must obtain in "the best of circumstances" as well, then, when no one tries to enrol more than once. So doubling $M(n)$ would not seriously overestimate $W(n)$ in the case at hand. But to see why that is so we must return to the situation of the relations (1) and (2).

So suppose again that k persons have been enrolled and that X is the identifier of the next enrollee. The chance that X will match some or other of the k stored templates is $\Phi(k) = 1 - (1-\xi)^k$ by (1); and this is the probability that X will match *at least one* of the stored templates of course. Now for any integer $1 \leq q \leq k$ we can ask what the chance is that X will match *at least q* of the stored templates; let $\Phi_q(k)$ denote this probability. We have $\Phi(k) = \Phi_1(k)$ of course: and generally we get

$$\Phi_q(k) = 1 - \sum_{r=0}^{q-1} C_r^k \xi^r (1 - \xi)^{k-r} \qquad \text{...(4)}$$

where $C_r^{\,k} = k! \,/\, r!\, (k - r)!$ is the "binomial coefficient" which counts how many distinct subsets of size r there will be in a set with k elements. To obtain the equality note that $\xi^r (1 - \xi)^{k-r}$ calculates the chance of matches between X and any subset of r templates among the k stored templates. There are C_r^k distinct such subsets; so the event of X's matching exactly r stored templates may occur in any one of C_r^k mutually exclusive ways, whence $C_r^k \xi^r (1 - \xi)^{k-r}$ is the probability of exactly r matches between X and the k stored templates; and the sum of these terms on the right of (4) will equal the probability $1 - \Phi_q(k)$ of *fewer* than q matches then, which gives us what we need. Computing with (4) will be intractable, again, when k is very large; and for our purposes we need workable approximations $\Phi_q(k)$. With a little effort one can show that

$$k\xi(1 - \xi)/(1 - \xi + k\xi) \leq \Phi_1(k) \leq k\xi \qquad \text{...(5)}$$

$$(1 - \xi)^{k-q} \xi^q C_q^{\,k} \leq \Phi_q(k) \leq q\, \xi^q C_q^{\,k} \qquad \text{...(6)}$$

The upper bound in (5) requires elementary calculus, only, and the lower bound no more. But the specification of the latter requires some ingenuity; and we must thank Nico Temme of the CWI in the Netherlands for having calculated the lower bound here. We should note, though, that when both ξ and $k\xi$ are minuscule quantities—as they would be for any feasible scheme of biometric identification—and the ratio between the bounds is practically 1 in (5), then the probability $\Phi_1(k)$ can be safely approximated with $k\xi$ simply. For $q = 1$ the upper bounds in (5) and (6) agree: but the lower bound in (5) is tighter. For $q \geq 2$ the lower bound in (6) is less workable than the upper: but luckily we shall only need the upper bound. The relation (6) requires only elementary calculus as well: but one must proceed by taking the term on the right in (4) as the value, for the argument ξ, of the *Incomplete Beta Function* with parameters k and $(k - q + 1)$. The derivations and calculations for the relations (2) and (5) and (6) are set down in an accompanying technical supplement (available online along with the text of this paper). The Incomplete Beta Function would not be a familiar thing to scientists and engineers generally. But it is very much a useful tool to anyone assessing the accuracy of biometric devices, as the compilation [3] shows. One might well expect, then, that the UIDAI has some specialist adept at using the function: who would have been able to perform all the calculations carried out in [1] based on the results of experiments that were reported in [2].

We can now outline the calculations which give us the second column in our table. We need an estimate of ξ to begin: which we could obtain from (2) if we had an estimate of the probability $\Phi_1(k)$ for some suitably large k. We had mentioned two experiments conducted by the UIDAI. The first of those was to estimate this probability, and it was performed when 84 million persons had been enrolled. The experiment is reported in [2]. It consisted of 4 million trials of the following description: in each trial a different template is picked from the stored templates, and compared against the remainder to see if a match occurs. Assuming that no one had been enrolled more than once, as [2] in fact does, the number of matches yields an estimate of $\Phi_1(k)$ for $k = 84 \cdot 10^6$. The report of the experiment in [2] records 2309 matches from $4 \cdot 10^6$ trials conducted, and takes $\phi \equiv 2309/4 \cdot 10^6$ as the estimate we are seeking. We shall use it; and putting ϕ for $\Phi_1(k)$ in the relation (2) we get

$$\phi/k \leq \xi \leq [-\log(1 - \varphi)]/k = \sum_{i=1\,<}^{\infty} \phi^i/i \cdot k$$

We have $\phi/k = (0.687202381) \cdot 10^{-11}$ here. To bound the series on the right note that the quantity $a \equiv \phi^4/4k$ is less than $(1/3) \cdot 10^{-21}$, so the tail from the 4-th term on is bounded by $a/(1 - \phi) < 10^{-21}$; and as $[\phi/k + \phi^2/2k + \phi^3/3k]$ comes to $(0.687400801) \cdot 10^{-11}$ now, we get

$$(0.687202381) \cdot 10^{-11} \leq \xi \leq (0.6874008011) \cdot 10^{-11} \qquad \text{...(7)}$$

Let us reiterate: the UIDAI could have estimated ξ here, the probability of a false positive for the composite biometric identifier they are using, by putting into the elementary relation (2) the result of their own experiment.

To proceed let $Y_1, Y_2, \ldots, Y_n$ be any listing of the biometric identifiers of the n distinct enrolled persons: in the order of their enrolment say. The comparison of each template to every other can be performed serially: for each $1 \leq k < n$ we compare Y_{k+1} to its k predecessors. Then $\Phi_1(k)$ is the chance that a match will occur with at least one of these predecessors. Assuming independent occurrence, as is usual, the total number of such templates can be estimated as

$$T_1(n) \approx \sum_{k-1}^{n-1} \Phi_1(k) \qquad \text{...(8)}$$

The sum on the right is the expected value of $(n - 1)$ independent Bernoulli trials, where for $1 \leq k < n$ the probability of success on the k-th trial is $\Phi_1(k)$ precisely. The k-th trial is a success if Y_{k+1} matches at least one of its k predecessors: so the sum will estimate the total number $T_1(n)$ of such matching templates.

Taking $\Phi_1(k) \approx k\xi$ yields $T_1(n) \approx \sum^{n-1}{}_{k-1} k = \xi[n(n-1)/2] = M(n)$ again, and this approximation seems safe enough given that $n < 2 \cdot 10^9$ in Table 1, for with the bounds on ξ in (7) the ratio $(1-\xi)/(1-\xi+k\xi)$ of lower to upper bound in (5) always lies between $(1-\xi) \approx 10^{11}/(10^{11}-1)$ and 1 now. For more precision one could get bounds on $T_1(n)$ by using (5) and (7) to get bounds on $\Phi_1(k)$. The knotty calculation that would involve was carried out in [1]. The difference proves negligible, though, for ξ here and the values of n in our table. We shall in a moment list the estimates of $T_1(n)$ thus obtained using the lower bound for ξ in (7): but to estimate the numbers $W(n)$ of duplicands we must count templates Y_{k+1} which match more than one of their predecessors. For $1 \leq q < n$ let $T_q(n)$ be this count: assuming independent occurrence again we may estimate it as

$$T_q(n) \approx \sum^{n-1}_{k-1} \Phi_q(k) \qquad \ldots(9)$$

The sum here is the expected value of $(n-1)$ independent Bernoulli trials, once more, where the k-th trial is a success if Y_{k+1} matches q or more of its predecessors, and the chance of success on the k-th trial is $\Phi_q(k)$ now. For our purposes it suffices to get an upper bound on $T_q(n)$ when $q \geq 2$: for which we shall use the upper bounds in (6) and (7) on $\Phi_q(k)$ and on ξ respectively. The totals $T_q(n)$ prove negligibly small for $q \geq 5$ here; and a routine calculation shows that with ξ and n as they are here we have

$$T_q(n) \approx q\,\xi^q n^{q+1}/(q+1)! \qquad \ldots(10)$$

for $2 \leq q \leq 4$. The calculation is in the same place where the relations (2) and (5) and (6) are derived. Estimating in the manner specified we get the numbers in Table 2:

Table 2

n	$T_1(n)$	$T_2(n)$	$T_3(n)$	$T_4(n)$	$T_5(n)$	$T_6(n)$
10^9	3436011	15751	41	0	0	0
$(1.1)10^9$	4157573	20964	59	0	0	0
$(1.2)\ 10^9$	4947856	27217	84	0	0	0
$(1.3)\ 10^9$	5806859	34604	116	0	0	0
$(1.4)\ 10^9$	6734582	43220	156	0	0	0
$(1.5)\ 10^9$	7731026	53158	206	1	0	0

The estimates of $T_1(n)$ are lower bounds: while for $2 \leq q \leq 6$ the estimates of $T_q(n)$ are upper bounds, having been obtained with the relation (10). The templates counted in $T_{q+1}(n)$ have already been counted in $T_q(n)$ of course, for a template that matches $(q+1)$ others certainly matches at least q others. Now a template that is counted in $T_q(n)$ will match at least q among the templates preceding it. But subtracting $T_{q+1}(n)$ from $T_q(n)$ counts the templates that match *exactly* q predecessors: and hence involve exactly $(q+1)$ templates. Let $R_q(n)$ be the set of templates which each match *exactly* q predecessors. To proceed we need an upper bound $U(n)$ on the number of templates that match some other. That cannot exceed twice the number of total matches now, assuming even that each and every match involves its own pair of templates: so from (3) we may set $U(n) = \xi n(n-1)$.

Let Y be a template in $R_1(n)$ and Z the unique predecessor it matches. Now $[U(n)-1]\xi$ is the chance that Z will match some *other* template besides Y : so $1 - [U(n)-1]\xi$ is the probability that Z will match none other besides Y. We must also consider that Y might match some successor. It is reasonable to assume that the $U(n)$ possibly matching templates will be uniformly distributed among the templates in the given listing; and reasonable to assume, as well, that the templates in $R_1(n)$ will be uniformly distributed among these $U(n)$ templates. So $[U(n) - 1]/2$ may be taken as the *expected number of succesors* that Y will have: whence $(1 - [U(n)-1]\xi/2)$ is the probability that Y will match none of these successors. The probability that any Y in $R_1(n)$ and its predecessor Z will form their own distinct matching pair comes to

$$\mu_1(n) \equiv (1 - [U(n)-1]\xi) \cdot (1 - [U(n)-1]\xi/2)$$

then: and we must count *twice* the difference $T_1(n) - T_2(n)$ multiplied by this uniform probability $\mu_1(n)$ in the number of duplicands $W(n)$ now, for a given Y in $R_1(n)$ or its matching predecessor Z might happen to match *some other* template, and multiplying by $\mu_1(n)$ corrects for the possibility of counting either Y or Z more than once in $W(n)$.

The same considerations apply to templates in a general $R_q(n)$ as well. Given a Y there let $[Y]$ be the set consisting of Y and its q matching predecessors: we must assess the probability that the elements of $[Y]$ form their own distinct set of $q+1$ templates, each matching some other among themselves, only, and none matching any other besides. That probability will come to

$$\mu_q(n) \equiv (1 - [U(n)-q]\xi) \cdot (1 - [U(n)-q]\xi/2) \qquad \ldots(11)$$

Now, the first factor in (11) computes the chance that a matching predecessor to Y matches none but Y and another of its q matching predecessors, if it matches any other template at all, and the second factor computes the chance that Y itself matches none besides these q predecessors. One readily sees that $\mu_q(n)$ is always small enough to serve as a probability here: for any $\xi < 10^{-11}$ and any $n \leq (1.5)10^9$ we have

$$[U(n)-q]\xi < U(n)\xi < \xi^2 n^2 < 3 \cdot 10^{-4} < 10^{-3}.$$

To our count of $W(n)$ we must add $(q+1)[T_q(n) - T_{q+1}(n)]\,\mu_q(n)$ for each applicable q then—multiplication by $\mu_q(n)$ correcting for the possibility, again, of counting more than once an element of $[Y]$ that might match a template that is not in $[Y]$ — and as $T_q(n) = 0$ for $q \geq 5$ here we may set

$$W(n) \approx \sum^4_{q-1} (q+1)\,[T_q(n) - T_{q+1}(n)]\,\mu_q(n) \qquad \ldots(12)$$

as our approximation, finally, of the count of duplicands. The numbers recorded in our first table were obtained by applying (12) to the estimates in the rows of our second table. The estimates of duplicands obtained thus may be taken as lower bounds for the actual numbers: because we have used the lower bound on ξ in (7) to estimate $T_1(n)$ always, while to estimate $T_q(n)$ for $2 \leq q \leq 4$, from the relation (10), we have used the upper bound on ξ in (7) and, as well, the upper bound on $\Phi_q(k)$ got from that using (6).

We noted that our estimate of $T_1(n)$ was tight, the bounds on the probabilities $\Phi_1(k)$ in (5) being very close. The same cannot be said for the bounds on $\Phi_q(k)$ in (6) for $q \geq 2$ though; and a referee had asked if our estimates were vulnerable for that reason. The prudent course would be to secure the principal contention of the paper by exhibiting lower bounds for $W(n)$ which are not contestable; and we shall do so by using the first and dominating term of the sum in (12) only, now, for we have

$$W(n) \geq 2[T_1(n) - T_2(n)]\,\mu_1(n) \qquad \ldots(13)$$

Certainly. Let us write $W_1(n)$ for the term on the right here: which undercounts the duplicands, note, because the counts $T_2(n)$ are not negligible while the counts $T_3(n)$ are, and we are ignoring the contribution $3[T_2(n) - T_3(n)]\,\mu_2(n)$ here. We shall use the lower bound for ξ in (7) for $T_1(n)$ and the upper for $T_2(n)$ again, as before, to estimate $W_1(n)$ also: and we get Table 3.

Table 3

n	$T_1(n)$	$T_2(n)$	$W(n)$	$W(n)/n$	$W_1(n)$	$W_1(n)/n$
10^9	3436011	15751	6887324	1/145	6840035	1/146
$(1.1)\ 10^9$	4157573	20964	8335451	1/132	8272509	1/133
$(1.2)\ 10^9$	4947856	27217	9921995	1/121	9840274	1/122
$(1.3)\ 10^9$	5806859	34604	11647035	1/112	11500000	1/113
$(1.4)\ 10^9$	6734582	43220	13510651	1/104	13400000	1/105
$(1.5)\ 10^9$	7731026	53158	15512927	1/97	15400000	1/98

The estimates of $T_1(n)$ and $T_2(n)$ in Table 3 are the same as in Table 2, of course, and the estimates of $W(n)$ are the same as in Table 1. Each entry under $W_1(n)$ is appreciably smaller, as we expect, than the corresponding one under $W(n)$. The ratios under $W(n)/n$ are the estimated proportions of duplicands from the first table: but the corresponding ratios under $W1(n)/n$ are very marginally smaller only, so we need not weaken our contention that the proportions of duplicands are too high.

Conclusions

We have considered the biometric identification programme of the UIDAI, and for varying levels of population estimated the proportion of duplicands: persons whose biometric identifiers match that of some other person. These proportions are too high: and indicate that the programme would badly fail to uniquely identify individuals. The estimation depends on the results of one experiment conducted by the UIDAI itself, and requires the elementary knowledge of the differential calculus, only, that any student of science or engineering would possess, and some acquaintance besides with one special function particularly relevant to assessing the accuracy of biometric identifiers. The experiment was performed in the very early stages of the programme, and the UIDAI should have been able even then to estimate the proportions of duplicands as we have here.

REFERENCES

[1] Mathews, V Hans (2013): "Biometric Identification: Device Specification and Actual Performance," considered for the operations of the Unique Identity Authority of India in Chapter 10 of *Advances in Biometrics for Secure Human Authentication and Recognition*, Dakshina Kisku, Phalguni Gupta, Jamuna Kanta Sing (eds), CRC Press, Taylor & Francis.

[2] UIDAI (2012): "The Role of Biometric Technology in Aadhaar Enrollment."

[3] Wayman, James (ed) (2000): US National Biometric Test Center, *Collected Works 1997–2000*, San Jose State University.

This is a technical supplement to the paper "Flaws in the UIDAI Process". The notation of the paper has been retained for the convenience of readers: but the arguments here are self-contained, and this supplement may be read independently.

1 Let ξ be any positive real number lying between 0 and 1. Set $\Phi(n) \equiv 1-(1-\xi)^n$ for any positive integer n. Our object is to derive the following inequalities:

$$(1) \qquad \Phi(n)/n \;\le\; \xi \;\le\; -\log[1-\Phi(n)]/n$$

$$(2) \qquad n\xi(1-\xi)/(1-\xi+n\xi) \;\le\; \Phi(n) \;\le\; n\xi$$

$$(3) \qquad (1-\xi)^{n-q}\xi^q\binom{n}{q} \;\le\; 1 \,-\, \textstyle\sum_{r=0}^{q-1}\binom{n}{r}\xi^r(1-\xi)^{n-r} \;\le\; q\xi^q\binom{n}{q}$$

The correspondences to the numbering of the paper are $(1) \leftrightarrow (2)$, $(2) \leftrightarrow (5)$ and $(3) \leftrightarrow (6)$. The central term in (3) would have been denoted $\Phi_q(n)$ in the paper: we shall do likewise: and $\Phi(n) = \Phi_1(n)$ of course. The binomial coefficients $\binom{m}{s}$ had been written as C^m_s in the paper.[1]

Set $g(t) = 1-(1-t)^n$ so that $g'(t) = n(1-t)^{n-1}$: it is elementary then that for $0<\xi<1$ we have

$$1-(1-\xi)^n \;=\; \left[n\int_0^\xi (1-t)^{n-1}\,dt\right] \;\le\; n\int_0^\xi dt \;=\; n\xi$$

since $0 < (1-t) \le 1$ when $0 \le t \le \xi < 1$. This inequality provides the lower bound in (1) and the upper bound in (2) already. To obtain the upper bound in (1) note first that $1-\Phi(n) = (1-\xi)^n$ by definition, which gives $\log[1-\Phi(n)] = n\cdot\log(1-\xi)$; so we must relate ξ and $\log(1-\xi)$ to proceed. For $0<x<1$ generally we have

$$\log(1-x) \;=\; -\int\frac{dx}{1-x} \;=\; -\int(1+x+x^2+\,\ldots)\,dx \;=\; -\left(x+\frac{x^2}{2}+\frac{x^3}{3}+\,\ldots\right)$$

since the series $\sum_{r=0}^{\infty} x^r$ converges absolutely. It is immediate that $\log(1-x) < -x$ then: and hence $x < -\log(1-x)$. So we have

$$\xi \;\le\; -\log(1-\xi) \;=\; \frac{-\log[1-\Phi(n)]}{n}$$

yielding the upper bound for (1). We get a lower bound on $\log(1-x)$ from its expression as the series above: simply note that

$$(x+x^2/2+x^3/3+\,\ldots) \;<\; (x+x^2+x^3+\,..) \;=\; x(1+x+x^2\ldots) = x/(1-x)$$

which gives us $-x/(1-x) < \log(1-x)$: and the bounds $x \;<\; -\log(1-x) \;<\; x/(1-x)$ will help obtain the lower bound in (2). For $0<x<1$ and positive integers n we now have

$$\begin{aligned} -n\log(1-x) \;&<\; nx/(1-x)\;;\quad 1-n\log(1-x) \;<\; (1-x+nx)/(1-x) \\ nx \;&<\; -n\log(1-x) \end{aligned}$$

from these bounds on $-\log(1-x)$; and these together yield

$$(1.1) \qquad nx\cdot\left[\frac{1-x}{1-x+nx}\right] \;<\; \frac{-n\log(1-x)}{1-n\log(1-x)}$$

Let $y>0$ next; to proceed we must take a detour and note that from $1+y \;<\; e^y$ we get

$$\begin{aligned} e^{-y} \,+\, ye^{-y} \;&<\; e^ye^{-y} \;=\; 1 \\ 0 \;&<\; 1 \,-\, e^{-y} \,-\, y\cdot e^{-y} \\ y \;&<\; 1 \,+\, y \,-\, e^{-y} \,-\, y\cdot e^{-y} \;=\; (1+y)\cdot(1-e^{-y}) \\ (1.2) \qquad \frac{y}{1+y} \;&<\; 1-e^{-y} \end{aligned}$$

[1] In the paper's equivalents of (1), (2) and (3) the letter 'k' appears where the the letter 'n' appears here: an innocuous change which should cause no confusion.

For $0 < x < 1$ set $t = -\log(1-x)$; we have $x = 1 - e^{-t}$ and $e^{-t} = 1 - x$ then, whence $e^{-nt} = (1-x)^n$ and $nt = -n\log(1-x)$; then for positive integers n we get

$$nx \cdot \left[\frac{1-x}{1-x+nx}\right] \; < \; \frac{-n\log(1-x)}{1-n\log(1-x)} \; = \; \frac{nt}{1+nt} \; < \; 1 - e^{-nt} \; = 1 \; - \; (1-x)^n$$

from (1.1) and (1.2) just above, as we need for the lower bound in (2). For the upper bound we have in (3) we must consider the *Incomplete Beta Function.* Set

$$\mathcal{B}_x(a,b) \; \equiv \; \int_0^x t^{a-1}(1-t)^{b-1}dt$$

for arguments a, b and any $0 < x \leq 1$ first; then $\mathcal{I}_x(a,b) \equiv \mathcal{B}_x(a,b)/\mathcal{B}_1(a,b)$ defines the Incomplete Beta Function for these arguments. It is usual to write $\mathcal{B}_1(a,b)$ as $\mathcal{B}(a,b)$ simply. Elementary integration, by parts, will give us

$$(1.3) \qquad \int_0^x t^j(1-t)^{n-j-1}dt \; = \; \frac{-t^j(1-t)^{n-j}}{n-j}\Bigg|_0^x \; + \; \left[\frac{j}{n-j}\right] \cdot \int_0^x t^{j-1}(1-t)^{n-j}dt$$

Setting $x = 1$ here yields the relation

$$(1.4) \qquad \mathcal{B}(j+1, n-j) \; = \; [j/(n-j)] \cdot \mathcal{B}(j, n-j+1)\,;$$

we have $\mathcal{B}(1,n) = \displaystyle\int_0^1 (1-t)^{n-1}dt \; = \; \frac{-(1-t)^n}{n}\Bigg|_0^1 \; = \; \frac{1}{n}$ to begin with; so iterating (1.4) gives

$$(1.5) \quad \mathcal{B}(j+1, n-j) \; = \; \frac{j \cdot (j-1) \cdot \, \ldots \, \cdot 2 \cdot 1 \cdot 1}{(n-j) \cdot (n-(j-1)) \cdot \, \ldots \, \cdot (n-2) \cdot (n-1) \cdot n} \; = \; \left[(n-j) \cdot \binom{n}{j}\right]^{-1}$$

Write $\mathcal{B}(a,b)$ as $\mathcal{B}_a^b$ to save space. We have $\mathcal{B}_{j+1}^{n-j} \cdot \mathcal{I}_x(j+1, n-j)$ on the left of equation (1.3) now, and $\mathcal{B}_j^{n-j+1} \cdot \mathcal{I}_x(j, n-j+1)$ for the integral on its right. By evaluating the first term on the right we obtain

$$\begin{aligned} \mathcal{B}_{j+1}^{n-j} \cdot \mathcal{I}_x(j+1, n-j) \; &= \; \frac{-x^j(1-x)^{n-j}}{n-j} \; + \; \left[\frac{j \cdot \mathcal{B}_j^{n-j+1}}{n-j}\right] \cdot \mathcal{I}_x(j, n-j+1) \\ &= \; \mathcal{B}_{j+1}^{n-j} \cdot \left[-\binom{n}{j}x^j(1-x)^{n-j} \; + \; \mathcal{I}_x(j, n-j+1)\right] \end{aligned}$$

because we have $1/(n-j) = \binom{n}{j} \cdot \mathcal{B}_{j+1}^{n-j}$ and $[j/(n-j)] \cdot \mathcal{B}_j^{n-j+1} = \mathcal{B}_{j+1}^{n-j}$ from (1.4) and (1.5) respectively. So

$$(1.6) \qquad \mathcal{I}_x(j+1, n-j) \; = \; \mathcal{I}_x(j, n-j+1) \; - \; \binom{n}{j}x^j(1-x)^{n-j}$$

now; it is immediate from the computation of $\mathcal{B}(1,n)$ that $\mathcal{I}_x(1,n) = 1 - [1-x]^n$; and then (1.6) will provide the inductive step for the equality

$$\mathcal{I}_x(q, n-q+1) \; = \; 1 \; - \; \sum_{j=0}^{q-1} \binom{n}{j}\xi^j \cdot (1-\xi)^{n-j}$$

To obtain the upper bound in (3) we need only note now that

$$\mathcal{B}_q^{n-q+1} \cdot \mathcal{I}_x(q, n-q+1) \; = \; \int_0^x t^{q-1}(1-t)^{n-q}\,dt \; \leq \; x^{q-1}\int_0^x (1-t)^{n-q}\,dt$$

generally; then, since $(1-t) \leq 1$ when $0 \leq t \leq \xi < 1$, as we have here, we finally get

$$\begin{aligned}
\mathcal{B}_q^{n-q+1} \cdot \mathcal{I}_\xi(q, n-q+1) &\leq \xi^{q-1} \int_0^\xi dt \leq \xi^q \\
\mathcal{I}_\xi(q, n-q+1) &\leq \xi^q / \mathcal{B}_q^{n-q+1} \\
&= \frac{\xi^q \cdot [n-(q-1)] \cdot [n-(q-2)] \cdots [n-1] \cdot n}{(q-1) \cdot (q-2) \cdot \cdots \cdot 2 \cdot 1} \qquad cf.\ (1.5) \\
&= \frac{\xi^q}{(q-1)!} \prod_{r=0}^{q-1} (n-r) \\
&= q\xi^q \binom{n}{q}
\end{aligned}$$

We do not need lower bounds on $\mathcal{I}_\xi(q, n-q+1)$ when $q > 1$: but we give them for completeness. As $1-t \geq 1-\xi$ for $t < \xi$ we have

$$\begin{aligned}
\mathcal{B}_q^{n-q+1} \cdot \mathcal{I}_\xi(q, n-q+1) &\geq (1-\xi)^{n-q} \int_0^\xi t^{q-1} dt = (1-\xi)^{n-q} \cdot \xi^q / q \\
\mathcal{I}_\xi(q, n-q+1) &\geq \frac{(1-\xi)^{n-q} \cdot \xi^q}{q \cdot \mathcal{B}_q^{n-q+1}} = \left[\frac{(1-\xi)^{n-q}}{q}\right] \cdot \left[\frac{\xi^q}{(q-1)!} \prod_{r=0}^{q-1} (n-r)\right] \\
&\geq (1-\xi)^{n-q} \xi^q \binom{n}{q}
\end{aligned}$$

This completes the derivation of the relation (3). The upper bound in (3) agrees with the upper bound in the relation (2) when $q = 1$: the latter bounds are a special case of the former. But the lower bound in (2) will exceed the lower in (3) unless

$$\begin{aligned}
(1-\xi)^{n-1} &\geq 1 - n \cdot \xi + \xi \geq (1-\xi)/(1-\xi + n \cdot \xi) \\
[1 - (n \cdot \xi - \xi)] \cdot [1 + n \cdot \xi - \xi] = 1 - (n-1)^2 \xi^2 &\geq 1 - \xi \\
(n-1)^2 \xi^2 &\leq \xi \\
(n-1)^2 \xi &\leq 1
\end{aligned}$$

This does not hold for the range of n in the paper and for the particular value of ξ there: but we have used the upper bound in (2) as a safe approximation of $\Phi_1(n)$.

2 We turn now to bounding the totals $T_q(N)$ in the paper: which was done for 6 values of N equally spaced between 1 billion to 1.5 billion. With $\Phi_q(n)$ as above for $1 \leq q \leq n$, and with N as the upper limit to the index n, we had $T_q(N) = \sum_{n=1}^{N-1} \Phi_q(n)$ there: and our object is to show that

$$T_q(N) \approx \frac{\xi^q N^{q+1}}{(q+1) \cdot (q-1)!}$$

for $2 \leq q \leq 4$, and the specified values of N, and the particular value $\xi = (0.6874008011)10^{-11}$ in the paper. It will prove very convenient to set $\lambda = 0.6874008011$ so that $\xi = \lambda \cdot 10^{-11}$ now. We may bound $\Phi_q(n) \equiv 1 - \sum_{r=0}^{q-1} \binom{n}{r} \cdot \xi^r \cdot (1-\xi)^{n-r}$ with $q\xi^q \binom{n}{q}$ using (3) above: upon which we have

$$\begin{aligned}
T_2(N) &\leq \sum_{n=1}^{N-1} 2\xi^2 \binom{n}{2} = \xi^2 \left[\sum_{n=1}^{N-1} n(n-1)\right] = \xi^2 \left[\sum_{n=1}^{N-1} n^2 - \sum_{n=1}^{N-1} n\right] \\
&\approx \frac{\xi^2 \cdot N \cdot (N-1) \cdot (2N-1)}{6} \\
&\approx \frac{\xi^2 \cdot 2 \cdot N^3}{6} = \frac{\xi^2 \cdot N^3}{3} = \frac{\xi^2 \cdot N^{2+1}}{(2+1) \cdot (2-1)!}
\end{aligned}$$

In going from line 1 to line 2 above we discard $\lambda^2 10^{-22} \sum_n n < N^2/10^{22}$ which is less than 10^{-3} for $N \leq (1.5)10^9$: and in going from line 2 to line 3 we again discard 3 summands where

N has power at most 2. We may do so because $T_2(N)$ is an integer. Continuing, we have

$$\begin{aligned}
T_3(N) \;&\leq\; \sum_{n=1}^{N-1} 3\xi^3\binom{n}{3} \;=\; \frac{\xi^3}{2}\left[\sum_{n=1}^{N-1} n(n-1)(n-2)\right] \\
&=\; \frac{\xi^3}{2}\left[\sum_{n=1}^{N-1} n^3 \;-\; 3\sum_{n=1}^{N-1} n^2 \;+\; 2\sum_{n=1}^{N-1} n\right] \\
&\approx\; \frac{\lambda^3}{2\cdot 10^{33}}\left[\sum_{n=1}^{N-1} n^3\right] \;=\; \frac{\lambda^3}{2\cdot 10^{33}}\left[\sum_{n=1}^{N-1} n\right]^2 \;=\; \frac{\lambda^3}{2\cdot 10^{33}}\left[\frac{N^4 \;-\; 2N^3 \;+\; N^2}{4}\right] \\
&\approx\; \frac{\lambda^3\cdot N^4}{4\cdot 2\cdot 10^{33}} \;=\; \frac{\xi^3\cdot N^{3+1}}{(3+1)\cdot(3-1)!}
\end{aligned}$$

In going from line 2 to line 3 we discard summands where the power of N is 3 or less, for these will not exceed 10^{-5} in absolute value: and we discard terms for the same reasons in going from line 3 to line 4. Going on in this manner we obtain

$$\begin{aligned}
T_4(N) \;&\leq\; \sum_{n=1}^{N-1} 4\xi^4\binom{n}{4} \;=\; \frac{\lambda^4}{10^{44}\cdot 3!}\left[\sum_{n=1}^{N-1} n(n-1)(n-2)(n-3)\right] \\
&\approx\; \frac{\lambda^4}{10^{44}\cdot 3!}\left[\sum_{n=1}^{N-1} n^4\right] \\
&\approx\; \frac{\lambda^4}{10^{44}\cdot 3!}\left[\frac{N^5}{5} - 10\left(\sum_{n=1}^{N-1}(n^3+n^2)\right) - 5\left(\sum_{n=1}^{N-1} n\right) - N\right] \\
&\approx\; \frac{\lambda^4 N^5}{5\cdot 3!\cdot 10^{44}} \;=\; \frac{\xi^4 N^{4+1}}{(4+1)\cdot(4-1)!}
\end{aligned}$$

and the rationale for discarding terms above should be clear now from what has already been said. Proceeding in this fashion will yield

$$T_5(N) \;\approx\; \frac{\lambda^5 N^6}{6\cdot 4!\cdot 10^{55}} \;<\; \frac{1}{3\cdot 4!\cdot 10}$$

for $N \leq (1.5)10^9$ and $\lambda < 0.69$: a quantity we may round to 0 since, to note it again, $T_q(N)$ is always an integer: and we need go no further now, since $T_r(N)$ only decreases as r increases.

Hans Varghese Mathews, Centre for Internet and Society