The UIDAI is bound to fail

A legal challenge is being mounted in the Supreme Court, currently, to the program of biometric identification that the Unique Identity Authority of India, the UIDAI henceforth, 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: the conclusion of which our title advertises already. We shall be using data that was available to the UIDAI: and shall employ elementary ways of calculating only. It should be recorded immediately that a technical paper by the present writer [1] has been of some use to the plantiffs: to which reference will be made 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 finger-prints and iris-images to generate a combination of biometrics for each individual: and the essay that follows 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 that results itself, 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 the table below lists the number of duplicands to be expected even in such circumstances, and *conservatively* expected, when the population will almost all have been identified.

identified, in billions	**duplicands** expected	**duplicands/identified**
1.0	6,824,248	1/147
1.1	8,251,493	1/133
1.2	9,812,984	1/122
1.3	11,508,424	1/113
1.4	13,337,513	1/105
1.5	15,299,957	1/98

The last column 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 122 persons to be a duplicand. The last row discloses that, by the time the population reaches 1.5 billion, they should expect one person in every hundred 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 should know. One cannot believe that there was no one with the requisite ability, at that time, among the personnel of the UIDAI: considering especially that the organization was then headed by a man who had been awarded a Padma Bhushan for his contributions to Science and Technology. One can only conclude that the individuals in charge of the UIDAI were either reckless or negligent: reckless if they calculated the embarassing numbers but went ahead nonetheless with a very poor biometric identification scheme: and negligent if they did not do the calculations.

A biometric is a numerical representation of an organic feature we said: which we shall regard as the *output* of some device upon its 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: 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: and 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 biometrized are presumably such as to warrant the assumption.

Let ξ be the probability of a false match 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 have biometrcially identified some k different individuals: whose identifiers $Y_1, Y_2, \ldots, Y_k$ are stored in some database $\mathbf{D}$ say. It is usual to call these stored identifers *templates*: and to say that the organic sources identified thus have been *enrolled*: so in our table above 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. In the best of circumstances the sources $SX, SY_1, SY_2, \ldots, SY_k$ are distinct persons: so ξ is uniformly the chance a match $X \sim Y_i$ between X and any stored template Y_i in $\mathbf{D}$: 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: and let us suppose so for the moment as well, for simplicity, deferring the question to an appendix. The chance that X will match *none* of the n 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 the relation

$$1 - \Phi(k) \;=\; (1 - \xi)^k \tag{1}$$

The probability $\Phi(k)$ can be reliably estimated when n 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

$$\Phi(k)/k \;\leq\; \xi \;\leq\; -\log[1 - \Phi(k)]/k \tag{2}$$

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.

Suppose that a total of n individuals are to be enrolled: and we must take $n \geq 2$ here of course. Each template will have been compared to every other when the enrollment is complete: and $n(n-1)/2$ comparisons will have been made, now, between the elements 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

$$M(n) \;=\; \xi[n(n-1)/2] \tag{3}$$

now for the expected number of matches when a total of n different individuals have been enrolled. Our table does not list the expected number of matches of course, 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: our table 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: 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 enroll more than once: which had been estimated as 0.0005 by the UIDAI in their report [2]: and the result (†) must obtain in "the best of circumstances" as well, then, when no one tries to enroll 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: and let $\Phi_q(k)$ denote this probability. We have $\Phi(k) = \Phi_1(k)$ of course: and generally we get

$$\Phi_q(k) \;=\; 1 - \textstyle\sum_{r=0}^{q-1} C_r^k \xi^r (1-\xi)^{k-r} \tag{4}$$

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

$$k\xi(1-\xi)/(1-\xi+k\xi) \leq \Phi_1(k) \leq k\xi \tag{5}$$

$$(1-\xi)^{n-q}\xi^q C_q^k \leq \Phi_q(k) \leq q\xi^q C_q^k \tag{6}$$

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 Professor 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 miniscule 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 $(k-q+1)$. In the appendix we shall provide a link, for interested readers, to where the calculations for the relations (2) and (5) and (6) are set down. 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]: and we may particularly expect this because, to note it again, the UIDAI was headed at the time by a man who had received a Padma Bhushan for his contributions to Science and Technology.

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: 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

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

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

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

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 enrolled persons: in the order of their enrollment say. The comparison of each template to every other can performed serially: for each $1 \leq k < n$ we compare Y_{k+1} to its k predecessors: and $\Phi_1(k)$ 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 will be

$$(8) \qquad T_1(n) \;=\; \textstyle\sum_{k=1}^{n-1} \Phi_1(k)$$

Seting $\Phi_1(k) \approx k\xi$ yields $T_1(n) \approx \; \xi \sum_{k=1}^{n-1} k \; = \xi[n(n-1)/2] = M(n)$ again: and this approximation seems safe enough given that $n < 2 \cdot 10^9$ in our table: 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 and $(1-\xi) \approx 10^{-11}/(10^{11}-1)$ 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]: but the difference proves negligible for ξ here and the values n in our table. We shall in a moment list the counts $T_1(n)$ 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. So for $1 \leq q < n$ generally set

$$(9) \qquad T_q(n) \;=\; \textstyle\sum_{k=1}^{n-1} \Phi_q(k)$$

next. Assuming independent occurrence again the total $T_q(n)$ counts the number of templates that match at least q of their predecessors. 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 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

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

for $q \geq 2$. The calculation can be found in the same place where the derivations of the relations (2) and (5) and (6) are detailed. Counting in the manner specified we get the following table:

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 counts $T_1(n)$ are lower bounds: while for $2 \leq q \leq 6$ the counts $T_q(n)$ are upper bounds, having been obtained with the relation (10) just above. 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* $(q+1)$ preceding templates. The sum

$$\begin{aligned} T_+(n) &= 3[T_2(n) - T_3(n)] + 4[T_3(n) - T_4(n)] + 5[T_4(n) - T_5(n)] \\ &= 3T_2(n) + T_3(n) + T_4(n) \end{aligned}$$

counts the templates that match more than one predecessor: and subtracting $T_+(n)$ from $T_1(n)$ yields the count of templates which each match exactly one preceding template. Let $\mathbf{R}_1(n)$ be this set of templates which each match exactly one predecessor. To proceed we need an upper bound $U(n)$ on the number of templates match some other: and that cannot exceed twice the number of total matches now, assuming that each and every match involves its own pair of templates: so from (3) we may set $U(n) = \xi n(n-1)$.

Suppose Y is a template in $\mathbf{R}_1(n)$ next. Let Z be the unique predecessor that Y matches. Now $[U(n)-1]\xi$ is the chance that Z will match some *other* template besides Y : and $1 - [U(n) - 1]\xi$ then the probability that Z will match none other besides Y. We must also consider the chance that Y might match some *successor*. It seems 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 $\mathbf{R}_1(n)$ 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. The probability that any Y in $\mathbf{R}_1(n)$ and its predecessor Z will form their own distinct matching pair comes to

$$\mu(n) \equiv (1 - [U(n) - 1]\xi)(1 - [U(n) - 1]\xi/2) \tag{11}$$

now: and the number of duplicands will equal or exceed *twice* the difference $T_1(n) - T_+(n)$ multiplied by this uniform probability $\mu(n)$ therefore. We must add the subtracted count $T_+(n)$ to that, of course, because the organic sources of the identifiers discounted from $T_1(n)$ thus will be duplicands also: and we will have

$$W(n) \;\geq\; 2[T_1(n) - T_+(n)] \cdot \mu(n) \;+\; T_+(n) \tag{12}$$

specifying a lower bound on the number of duplicands, finally, using the upper bound on ξ given by (7) again. The counts of duplicands in our first table were obtained by applying (12) to the rows of our second table.

Conclusion We have considered the biometric identification program 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 the program would badly fail to uniquely identify individuals. The estimation depends on the results of one experiment conducted by the UIDAI itself: requires the elementary knowledge of the differential calculus, only, that any student of science or engineering should possess: and some acquaintance besides with one special function particularly relevant to assessing the accuracy of biometric identifiers. The UIDAI should have among its personnel someone with the requisite ability: or should have had some person so able when the experiment was performed at least: for, to note it yet once more, the organization was then headed by man who had been awarded a Padma Bhushan for his contributions to Science and Technology. The experiment was performed in the very early stages of the program: and the UIDAI should have been able even then to estimate the proportions of duplicands as we have here. If the estimation was carried out the individuals in charge of the UIDAI have been reckless in proceeding with the program: and if not they have been grossly negligent.

APPENDIX

Suppose X, Y and Z are biometric identifiers deriving from distinct organic sources. Let ξ be the chance of a false positive. In the specification of the probabilities $\Phi_q(k)$ the following was implicitly assumed: suppose $Y \sim Z$ and $X \sim Y$: before X and Z are compared the probability of the match $X \sim Z$ is ξ still. The assumption is routine in assessing the accuracy of biometric identifiers: but it might be prudent to proceed otherwise. X is to be compared to $Y_1, Y_2, \ldots, Y_k$ suppose: with the organic sources $SX, SY_1, SY_2, \ldots, SY_k$ all being distinct again. The identifiers Y_i may be grouped into subsets whose elements mutually match: call these *matching classes.* An identifier which matches no other will be the sole member of its matching class: call such a class a *unit.* Suppose there are $k_0 \leq k$ matching classes in all. The relation

$$1 - (1-\xi)^{k_0} \;\leq\; \Phi_1(k) \;\leq\; 1 - (1-\xi)^k \tag{A1}$$

seems to be a safer way of estimating the chance that X will match at least one among the Y_i. The analogue of relation (2) above would then be

$$\Phi_1(k)/k \;\leq\; \xi \;\leq\; -\log[1 - \Phi_1(k)]/k_0 \tag{A2}$$

The lower bound here is the same as in (2) while the upper bound is larger: so the numerical bounds for ξ in (7) are usable values. From the experiment conducted by the UIDAI one might venture to estimate the number of matching classes there were

among $k = 84{\cdot}10^6$ identifiers. Given 2309 matches in $4{\cdot}10^6$ random trials there are at most 1155 matching classes that are not units, we may suppose, among any 4 million: each class would have two members now: and we might assume $12 \cdot 1155 = 13860$ such classes among the 84 million then. The other matching classes would be units: and we would get $k_0 = (84 \cdot 10^6 - 2 \cdot 13860) + 13860 = 84 \cdot 10^6 - \cdot 13860$ matching classes altogether. Repeating the calculations that gave us (7) we now have

$$\text{(A7)} \qquad (0.687202381) \cdot 10^{-11} \;\leq\; \xi \;\leq\; (0.687514195) \cdot 10^{-11}$$

The upper bound is marginally larger than it was before. For $q \geq 2$ the analogue of the relation (4) now is the inequality

$$\text{(A4)} \qquad \Phi_q(k) \;\leq\; 1 - \textstyle\sum_{r\,=\,0}^{q-1} C_r^k \xi^r (1-\xi)^{k-r}$$

But we only need upper bounds on $\Phi_q(k)$ to obtain, for $2 \leq q \leq n$, the upper bound on $T_q(n)$ that (10) provides. The calculation of $T_1(n)$ does not change since the lower bounds on ξ are the same in (7) and (A7). Repeating with the upper bound in (A7) the calculation of $T_2(n)$, $T_3(n)$, ... one finds that $T_q(n) \approx 0$ for n in our range and $q \geq 5$ again: and using the relation (11) we now get the following estimates of duplicands and their proportions:

identified, in billions	**duplicands** expected	**duplicands/identified**
1.0	6,824,233	1/147
1.1	8,251,472	1/133
1.2	9,812,957	1/122
1.3	11,508,388	1/113
1.4	13,337,471	1/105
1.5	15,299,903	1/98

The counts of duplicands are marginally smaller now, as we expect, but the rounded-off proportions remain the same: so we need not revise our large conclusions.

A link would be provided for interested readers, we had said, to details regarding certain relations we have used above: to the derivation of (2), (5), (6) and (10) specifically: and that is set down in the technical paper available at ⟨⟨ URL? ⟩⟩.

Hans Verghese Mathews

references

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

[2] *The Role of Biometric Technology in Aadhaar Enrollment*
Unique Identity Authority of India; online publication, 2012

[3] *Evaluation of Biometric Identification Systems*
William Barrett; Summary, for the US National Biometric Test Center, San Jose State University, San Jose, California, U.S.A.