A request for specifics

The author of a technical paper will be alarmed when he is convicted of "serious mathematical errors" by someone who has not bothered himself with "going too deep into the mathematics" used. The man must possess miraculous powers of divination one feels: fears rather. The UIDAI seems to have even such formidable diviners in their employ: who have dismissed just so peremptorily, in their rebuttal, the calculations made in my paper titled *Flaws in the UIDAI process.* The paper appeared in the issue of this journal dated to February 27 of this year. The rebuttal was published in the issue dated to the 12th of March. The interested reader can confirm that I have only repeated what was said there. The rebuttal does not specify, in any way, the *mathematical* mistakes I am supposed to have made. So I shall rehearse the relevant calculations very broadly: and the experts of the UIDAI will then exhibit, I trust, the specific mistakes they impute to me.[1]

I repeat here that the calculations in my paper pertain to "the best of circumstances" for the biometric identification of a population: when no one attempts, that is, to be "enrolled" more than once in the biometric database. The biometric identifiers stored in the database are usually called *templates.* Suppose that k different persons have been enrolled: what the UIDAI call *the false positive identification rate,* or the FPIR for short, is the probability that the biometric identifier of the *next* enrollee will match *at least one* of the stored templates. That probability is a function of the size k of the database. It was called $\Phi(k)$ in my paper, or $\Phi_1(k)$ equivalently, to emphasize that the new identifier must match at least one of the stored templates; and to prevent confusion the FPIR will always be called $\Phi_1(k)$ here.

Besides the FPIR one has the simple chance of a *false positive*: which the probability that the identifiers of different persons will match. That was called ξ in the paper. Note that ξ pertains to the biometric system *itself,* now, in contrast to the FPIR: which latter, to say it again, is a function of the number of stored templates. Here is the basic relation between these probabilities: when there are k stored templates we have

$$\text{(a)} \qquad 1 - \Phi_1(k) \;=\; (1-\xi)^k$$

The only assumption behind (a) is that the occurrences of matches between a new identifier and the stored templates, if any should take place, will all happen *independently* of each other. The assumption of independence is standard in the assessment of biometric systems: and if one refuses to countenance this fundamental assumption some reason must be provided. From the basic and fundamental relation (a) elementary calculus will yield the relation

$$\text{(b)} \qquad \Phi_1(k)/k \;\leq\; \xi \;\leq\; -\log[1-\Phi_1(k)]/k$$

[1] My reply to the UIDAIs attempted rebuttal was sent in to the EPW a few days after that appeared in print: and published as a "web exclusive" article in Volume 51, Issue Number 36 of the EPW, on 03/09/2016.

The relations (a) and (b) were numbered (1) and (2) in the paper. The derivation of (b) was detailed in the technical supplement which was made freely available online. The commentators on my paper have looked at that I trust: and I shall be only too glad if they can detect any mistakes there. So if one has a reliable estimate of $\Phi_1(m)$ now, for some m or other, then one can estimate ξ with that: and the estimate of ξ will be *as reliable* as the estimate of $\Phi_1(m)$ here. An experiment performed by the UIDAI allowed them to estimate $\Phi_1(m)$ for m equalling 84 million. They have taken that as a reliable estimate. So they are bound to regard the bounds on ξ that I derived therefrom, in equation (7) of the paper, as equally reliable —– unless the 'ideal' operations of the calculus are suspect, somehow, when they are performed upon numbers that purport to measure quantities in the 'real' world. But if the experts of the UIDAI have grave doubts regarding the mathematics that is fundamental to science and engineering, then they should declare themselves, really, for the profound sceptics they are.

We have seen so far that ξ can be reliably estimated from a reliable estimate of $\Phi_1(m)$ for some *one* value of m : and it turns out that one can estimate $\Phi_1(k)$ for *any other* positive integer k then, from the bounds given by

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

The relation (c) was numbered (5) in the paper: and its derivation too was detailed in the technical supplement. Only elementary calculus is required again. The rebuttal avers that "an extrapolation curve" for the FPIR "has to be developed empirically and cannot be derived mathematically." For a fixed ξ the bounds in (c) can be regarded as functions, of the number k of stored templates, which return bounds on the FPIR for a given k. If these are the "extrapolation curves" referred to — just such "functions or curves" as "any extrapolation needs," to phrase the matter as the rebuttal does — then to suppose that these functions "cannot be derived mathematically" is simply wrong: for I have done precisely that. To sustain their contention the experts of the UIDAI must exhibit specific mistakes made in the derivation. I await the demonstration: failing which the experts of the UIDAI must accept that the bounds on $\Phi_1(k)$ I obtain from (c) are as reliable, again, as the estimate of the FPIR they had obtained, themselves, when 84 million persons had been enrolled —– unless, again, they altogether doubt the mathematics fundamental to science and engineering.

In the paper I had used my lower bound on ξ and the resulting upper bound $k\xi$ on $\Phi_1(k)$ to estimate, for varying levels n of the total population, the number of times an enrollee would find his or her identifier matching at least one stored template: the identifier of a previous enrollee that is. That sum was called $T_1(n)$ in the paper: and it was estimated there as

$$\text{(d)} \qquad T_1(n) \;\approx\; \textstyle\sum_{1 \leq k < n} \Phi_1(k)$$

The relation (d) was numbered (8) in the paper: and the sum on the right may be regarded as the expected value of $(n-1)$ Bernoulli trials where the chance of

success on the k-th trial is $\Phi_1(k)$ precisely. That these trials are independent is a fundamental assumption we had noted: but (d) will hold regardless because the expected value of a sum of random variables is the sum of the individual expected values no matter how the variables are jointly distributed. Though I had provided what I thought were good reasons for doing so, it may have been imprudent to estimate $T_1(n)$ using the upper bound $k\xi$ on $\Phi_1(k)$ *only*: without some correction offered using the lower bound in (c) as well. I shall make good the omission momentarily. But it cannot be a serious *mathematical* error to have employed $k\xi$ just so: not unless the derivation of (c) from (a) is flawed. So to sustain their contention the experts of the UIDAI must exhibit, again, some or other mistake in that derivation: and I await, again, the demonstration of putative error.

Let me carry out the correction adverted to just now. Estimating $T_1(n)$ with (d) using $k\xi$ for $\Phi_1(k)$ is direct: since $\sum_{1 \leq k < n} k\xi = \xi n(n-1)/2$ of course. No such ready calculation seems available if the lower bound in (c) is used instead: and to be completely exact one would have to separately compute, for a given n, each of the $(n-1)$ quantities $k\xi(1-\xi)/(1-\xi+k\xi)$ for $1 \leq k < n$ here. I had computed for 6 levels of the total population n in my paper, going from 1 billion to 1.5 billion in steps of 100 million. So to be exact here 7.5 billion distinct quantities would have to be computed. I have opted for a shortcut.

The lower bound for ξ is $(0.687202381) \cdot 10^{-11}$ here. Set $\lambda = 0.687202381$ first. Now when $m = j \cdot 10^6$ exactly, for any $j \geq 0$, one has the very close approximation

$$\text{(e)} \qquad \xi(1-\xi)/(1-\xi+m\xi) \;\approx\; \lambda/(10^6 \cdot (10^5 + j\lambda))$$

The shortcut I take is the following: for each k lying between one millionth value and the next millonth value — for $(j-1) \cdot 10^6 < k \leq j \cdot 10^6$ here, that is to say — I employ the approximation

$$\Phi_1(k) \;\approx\; k\xi(1-\xi)/(1-\xi+k\xi) \;\approx\; k\lambda/(10^6(10^5 + j\lambda))$$

We are using the lower bound in (c) here: but we are underestimating the probabilities $\Phi_1(k)$ even more by keeping the factor $\xi(1-\xi)/(1-\xi+k\xi)$ at the lowest value that it can take for k in the specified range. That value is $\xi(1-\xi)/(1-\xi+j \cdot 10^6 \cdot \xi)$ of course, for k appears only in the denominator of the factor. Treat the j-th million of the enrollees as a batch now: and let N_j count all those among these whose identifiers will match at least one of the stored templates. Set $Q_l = l \cdot 10^6$ for ease of writing. The approximation

$$N_j \;=\; \textstyle\sum_{Q_{j-1} < k \leq Q_j} \Phi_1(k) \;\approx\; [\lambda/(10^6(10^5 + j\lambda))] \cdot \sum_{Q_{j-1} < k \leq Q_j} k$$

provides a lower bound on N_j then. The relation $\sum_{r < i \leq s} i = (s-r)(s+r+1)/2$ is got by subtracting $(1+2+\,..\,+r)$ from $(1+2+\,..\,+s)$. That directly yields the sum on the right above: and we get

(f) $\quad N_j \;\approx\; [\lambda(2j-1)10^6+1]/[2(10^5+j\lambda)]$

Suppose the total population n is some multiple of a million: take $n = K\cdot 10^6$ some $K > 0$ that is. As N_j estimates the matches expected for each j-th million we get a lower bound for the total $T_1(n)$ of expected matches from

(g) $\quad T_1(n) \;\approx\; [\sum_{1\leq j\leq K} N_j] \;-\; \Phi_1(n)$

The correction by subtracting $\Phi_1(n)$ is needed since that computes the chance of a match for the $(n+1)$-st enrollee. Using the lower bound in (c) for the FPIR then, and computing with the formulae (f) and (g) and the stated value of $\lambda = \xi\cdot 10^{11}$ now, for the levels n of total population which I had considered in my paper, one gets the following table analogous to Table 3 there:

n	$T_1(n)$	$T_2(n)$	$W(n)$	$W(n)/n$	$W_1(n)$	$W_1(n)/n$
10^9	3420339	15751	6855985	1/146	6808696	1/147
$(1.1)10^9$	4136726	20964	8293764	1/133	8230822	1/134
$(1.2)10^9$	4920805	27217	9867904	1/122	9786183	1/123
$(1.3)10^9$	5772486	34604	11578305	1/112	11474398	1/113
$(1.4)10^9$	6691674	43220	13424859	1/104	13295073	1/105
$(1.5)10^9$	7678280	53158	15407469	1/97	15247830	1/98

The counts $T_2(n)$ estimate, as before, the total number of times the identifier of an enrollee will match at least two stored templates. They are the same here as in the paper, having been computed again with upper bounds derived from the Incomplete Beta Function: the use of which the experts of the UIDAI do not impugn in their rebuttal. The other counts of matches and duplicands here are lower, of course, than in Table 3 of the paper: but the ratios $W(n)/n$ and $W_1(n)/n$ of duplicands to population are the same for the last three rows, note, and only very marginally smaller for the first three rows. The paper had explained why the last column gives the safer approximation. I shall set side by side the safer estimates $W_1(n)/n$ of duplicand ratios, called rU and rL below, that are obtained by using the upper and lower bounds for the FPIR, respectively, that the relation (c) provides:

n	rU	rL
10^9	1/146	1/147
$(1.1)10^9$	1/133	1/134
$(1.2)10^9$	1/122	1/123
$(1.3)10^9$	1/113	1/113
$(1.4)10^9$	1/105	1/105
$(1.5)10^9$	1/98	1/98

In the paper I had maintained that the difference would be negligible if one used the lower bound for the FPIR instead of the upper: because the ratio between these bounds in (c) is almost 1 for ξ and the range of the population here. I

can continue to do so: considering especially that, in taking the shortcut above, one is getting counts for $T_1(n)$ which are lower than what the lower bound in (c) would give. The safer approximation of duplicand ratios had been declared *incontestable* in the paper: I shall maintain the claim. To contest the ratios rU and rL above the UIDAI will have to exhibit the "serious mathematical errors" that I am supposed to have commited in estimating them: which I could have made only in *deducing,* using elementary calculus, the relations from which these ratios derive. I await the UIDAI's demonstration of my deductive errors.

The burden of UIDAI's rebuttal of my paper seems to be this: that I have made a mathematical *projection* of quantities that could only be *found,* or so the UIDAI had supposed, by empirical means. That a problem has been solved in a novel way is no argument against the proposed solution: and I must emphasize that my projections employ only the mathematics basic to science and engineering. The UIDAI may reject these projections only by exhibiting some deductive or computational error made in making them: or, if no such errors can be detected, only by denying the basic relation (a) —– or else the UIDAI may reject my projections only by casting grave doubt, altogether, on the propriety of using elementary calculus and probability to assess the efficacy of biometric systems. But perhaps the experts of the UIDAI — who are able to divine "serious mathematical errors," after all, "without going too deeply into the mathematics" —– perhaps those wizards are bold enough to do just that.

Hans Verghese Mathews, Bangalore, 24/03/2016

Addendum Summarizing their rebuttal the experts of the UIDAI assert that I have erred "on two counts" mainly. The first of these concerns how "false duplicates" are resolved by the UIDAI: but my paper does not address the issue at all. The second supposed error is that "the paper assumes an extrapolation function on a single datapoint, which is incorrect for multimodal biometric systems." For a fixed ξ the bounds in (c) can be regarded, we already noted, as functions of the number k of stored templates. But to complain that they are *extrapolations from one data point* — and are merely *assumed* moreover —– that is disingenuous and misleading: for these functions have not been "extrapolated" from any "datapoint" at all, to emphasize the circumstance once more, but have been *deduced* from the basic relation (a) rather. The relations (b) and (c) together *do* allow one to calculate the FPIR for different totals k of stored templates from the single "datapoint" of an FPIR estimated for some particular total m of stored templates: and one cannot insist that doing so is an error unless, to say it again, one can demonstrate how these relations themselves are erroneous. Coming to the biometric system itself now: the rebuttal does not say why or how an extrapolation illicit for a "multimodal" system could be licit for a "unimodal" one. But let me note, regardless, that the primary relations (a) and (b) and (c) would hold for any biometric system: and the considerations of my paper would apply equally to both "unimodal" and "multimodal" biometric systems.

Postscript, 03/09/2016 I reproduce below the UIDAI's response to my request for specifics regarding the "serious mathematical errors" I was supposed to have made:

> "The author has used linear extrapolation to project that 1/145 enrolments could result in false positive identification because of biometric de-duplication. There is sufficient evidence from biometric literature, as cited in our earlier rebuttal (EPW, March 12, 2016), to show that extrapolation using a single data point can lead to erroneous conclusions. This is especially true for a multi-modal biometric system such as Aadhaar. Most importantly, it has been admitted by the author that operational resolution of false positives is not of interest to him. The operation resolution of UIDAI process ensures that no person is denied of Aadhaar on account of false positives. Since the author chooses to not look at the operational resolutions, he cannot justify the title of his paper Flaws in UIDAI process, since operational resolution is a part of the UIDAI process."

This response by the UIDAI was published in the EPW together with my request for specifics: and a letter by my colleague Pranesh Prakash was published alongside, which pointed out that "in fact, in their paper *Role of Biometric Technology in Aadhaar Enrollment,* the UIDAI states: *FPIR rate grows linearly with the database size*." The UIDAI itself had made of use the upper bound in (c), that is to say, to project how the false positive identification rate would grow as enrollment proceeded: and I do not think anything more need be said, now, regarding how competent the personnel of the UIDAI would be in assessing the efficacy of their biometric system.