Prisoners, Boxes, and Cards Riddle

Imagine we have two million prisoners, each numbered 1 to 2,000,000, a big room with two million boxes numbered 1 to 2,000,000, and two million cards, also numbered 1 to 2,000,000. The warden shuffles the cards and places one inside each box. He then explains to the prisoners the game he is setting up: one by one, the prisoners will be allowed to enter the room with the boxes, and each prisoner will be allowed to open one million (or more relevantly: *half*) of the boxes to see the cards inside. Each prisoner gets to choose which boxes he opens, but he cannot move any of the cards, and he must close the boxes again when he is done. Afterwards, he will exit the room and will be unable to communicate with any of his fellow prisoners. If every single prisoner finds the card that matches his inmate number, then the warden will set all of them free. They are allowed to discuss a strategy before the game begins, but once it has started there will be no more communication between prisoners.

If the prisoners decide to just open the boxes at random, then there is a 1 in $2^{2,000,000}$ chance that they will be set free. Imagine our entire observable universe is packed full of hydrogen atoms. And when I say "packed", I mean *packed*— imagine a ball of solid hydrogen packed as densely as possible— say, Hexagonal Close Packed, where the distance between adjacent nuclei is the same as the documented "diameter" of the hydrogen atom, *i.e.* the atoms are virtually touching. This gives a packing density of about 74%. Now imagine this ball is the size of the observable universe. Then imagine each of the hydrogen atoms is magically turned into a universe the same size as ours, each of which can then be packed with new hydrogen atoms. Call the process of turning all existing hydrogen atoms into universes and filling all the new universes with more hydrogen atoms a 'Splosion. After 5,442 'Splosions, take all the hydrogen atoms in all of the most recent generation of universes and turn them into hats, with each hat containing 70,000,000,000,000,000,000,000,000 red marbles (that's seventy septillion). I pick a single hat, then pick a single marble from that one hat, and I paint it blue. If you picked a random marble from a random hat, you'd have slightly better odds of picking the blue marble than the prisoners would have of going free.

Let's say instead of opening boxes randomly, the prisoners follow this strategy: have each prisoner first open the box whose number matches his own inmate number. It will contain a card with a number. He should then open the box whose number matches that card, which will contain a card with a different number. He'll open that number box next, and so on. He will stop once he either opens the box that contains the card that matches his inmate number or once he's opened one million boxes, whichever comes first. If every prisoner follows this strategy, their odds of freedom go to about 3 in 10.

I did not come up with this strategy, but upon hearing how insanely better this strategy supposedly is compared to picking boxes at random, I decided to pause Veritasium's video to figure out how this could possibly be. The riddle posed in the video used one hundred prisoners, boxes, and cards instead of two million, but I quickly realized that the odds of success with this strategy converges on a specific value as this number approaches infinity. So instead of 100 or 2,000,000 I simply used n to represent the number of prisoners/boxes/cards, with the intent of taking the limit as $n \to \infty$. I also wanted to know how the odds would change if the warden allowed each prisoner to open some fraction of the boxes *greater* than one-half, a fraction which I called q.

So here's the problem, rephrased in its bare-bones: we have n integers, 1 to n, arranged in a random order. The integers themselves represent the cards, and the place in which each integer sits represents the boxes. Take the 1^{st} integer in the sequence (*i.e.* the card that sits in box number 1), a, then go to the a^{th} integer, b, (*i.e.* the card that sits in box number a), then go to the b^{th} integer in the sequence, and so on. At some point (possibly even immediately), you will arrive at the integer "1" which directs you back to the 1^{st} integer, thus you have identified a loop that this permutation generated. Go to the next integer that has

not been visited yet and repeat. You will end up with a collection of loops which can be identified with a particular partition of n (*i.e.* a group of integers that sum to n).

Let's look at an example. Say $n = 5$ and the random sequence is 3, 2, 4, 1, 5. We start with the 1st integer, 3, then go to the 3rd integer, 4, then to the 4th integer, 1, which closes that loop of three integers. The next integer is 2, and the 2nd integer is 2, so that closes a loop of one integer. The only integer left is 5, which is in the 5th position, thus making another loop of one integer. This brings us to the issue of how to notate a specific partition. When $n \leq 10$, I think it's safe, easy, and efficient to simply list the loops in descending-size order. For this example, that would simply be 311. However, if $n > 10$, this notation gets very ambiguous (and potentially very expansive). Keeping with the convention of descending by loop size, we list the largest loop size followed by the number of occurrences of that size loop followed by the next largest loop size, etc. So using this notation, our example would be 3.1;1.2.

While we're on the subject of notation, let's quickly define a permutation count function, $\chi(x)$, as the number of unique permutations of n integers that generate a specific partition, x. It's important to note that the input for the function $\chi()$ is not a number, it is a specific partition. We'll define $\rho(x_i)$ as the ratio of unique permutations that generate one or more partitions to the total number of permutations ($n!$). For this function, the input is a *collection* of partitions. For example: x_i could refer to all partitions of 5 such that the largest part is 3, so x_1 would refer to partition 311 and x_2 would refer to partition 32, and $\rho(x_i) = \rho(311, 32) = 40/120$ (you'll just have to take my word on that for now, or you could try to find the 40 permutations yourself).

The last bit of notation to introduce is for the probability that a random permutation of n integers contains a loop of size $\geq q \cdot n$ where q is a rational number between 0 and 1. And that notation is this: $P_q(n)$.

Lemma:

$\chi(n.1) = (n-1)!$

Proof:

If we want to know how many permutations generate the partition $n.1$, then we take the number of options for the first integer times the number of options for the second integer, etc. For the first integer, every option is open to you except 1 because that would close a loop of size 1, so you have $n-1$ options. For the second integer, you can use any integer except for 2 and whatever integer occupies the first place, so that leaves $n-2$ options. But what if 2 is the first integer? Wouldn't you then have $n-1$ options again? No, because in this case, we'd be restricted from using 1 as the second integer because that would close a loop of size 2. For the third integer, you are restricted from using 3 and whatever the first two integers were. If 3 was in one of the first two spots, then you are restricted from using the place that contains 3, or if that integer was already used as well, the place that contains *that* integer. For example, if we have the permutation 2, 3, __, ..., we are clearly restricted from using 2 and 3 since those have already been used, but we are also restricted from using 1 (*i.e.* the place that contains the place that contains 3) because that would close a loop of size 3. That leaves $n-3$ options. It's easy to see that for the i^{th} integer there are $n-i$ options, therefore the number of permutations is $(n-1) \cdot (n-2) \cdot (n-3) \cdot \ ... \ \cdot 1 = (n-1)!$

Lemma:

$\rho(m.1; x_i) = 1/m$ where $m.1; x_i$ is the collection of all partitions for a given $m > 1/2\ n$

Proof:

First, let's observe that $\chi(m.1; x) = \binom{n}{m} \cdot (m-1)! \cdot \chi(x)$. This should make sense upon inspection, but let's walk through it anyway. The factor of $\binom{n}{m}$ is for our freedom to choose which places in our permutations of n are a part of the m loop. The factor of $(m-1)!$ is for the number of possible permutations of the integers in the m loop (see lemma above). The factor of $\chi(x)$ is for the number of possible permutations of the integers

that generate the x partition. We can simplify this equation a bit:

$$\begin{aligned}\chi(m.1;x) &= \binom{n}{m} \cdot (m-1)! \cdot \chi(x) \\ &= \frac{n!(m-1)!}{m!(n-m)!} \cdot \chi(x) \\ &= \frac{n!}{m(n-m)!} \cdot \chi(x)\end{aligned}$$

Let's now sum over all possible partitions x:

$$\begin{aligned}\sum_x \chi(m.1;x) &= \frac{n!}{m(n-m)!} \cdot \sum_x \chi(x) \\ &= \frac{n!}{m(n-m)!} \cdot (n-m)! \\ &= \frac{n!}{m}\end{aligned}$$

Remember, the total number of permutations of n integers is $n!$ and this is what we divide by to get $\rho(m.1;x_i)$. Therefore, $\rho(m.1;x_i) = 1/m$[†]

Conjecture:

$\lim_{n\to\infty} P_q(n) = \ln(1/q)$ for $q > 1/2$

Proof:

Because of the lemma above, we know that for $q > 1/2$,

$$P_q(n) = \sum_{t=\lceil q \cdot n \rceil}^{n} \frac{1}{t}$$

The first thing to note is that since we're taking the limit where n approaches infinity, we may as well drop the ceiling brackets. Here's why we can do that: for any given q (which is rational), there's an infinite number of evenly spaced n's such that $q \cdot n$ is an integer. If the limit exists, then it doesn't matter whether we use all n's or just those ones. So let's revise our equation a little. q can be represented as the fraction r/s where r and s are positive integers ($r < s$). If we only want to focus on n's such that $n \cdot r/s$ is an integer, then we simply choose all of the multiples of s to be our chosen n's. Then the equation can be written like this:

$$\lim_{n\to\infty} P_q(n) = \lim_{i\to\infty} \sum_{t=i\cdot r}^{i\cdot s} \frac{1}{t}$$

where i is a positive integer. Let's now think of the sum not as a sum of numbers, but as a sum of areas of adjacent rectangles whose widths are all 1 and whose heights are $1/t$. As i approaches infinity the heights of these rectangles (and therefore also their areas) approach zero, but the number of rectangles approaches infinity. This sounds like what an integral does, except the wrong side of the rectangles is shrinking! What if instead of keeping the width of each individual rectangle constant as we increase i, we scale them down

[†]The reason we stipulated $m > 1/2\ n$ was to ensure that no loop in x_i could be of size m. If we had to account for more than one loop of size m, our equations would have gotten messier. Anyway, this limitation fits very well with what we're ultimately proving.

by a factor of $1/i$? In order to maintain the area of each rectangle, we would have to scale their heights up by a factor of i. The combined width of all the rectangles, whereas before was $i(s-r)+1$, is now $s-r+\frac{1}{i}$, which, as i approaches infinity, approaches $s-r$.

If we're successful in turning this into an integral, the interval over which we integrate would clearly be r to s. Let's look at what happens to the heights of the rectangles. Before rescaling, each rectangle had a height of $1/(i \cdot r + j)$ where j is an integer that goes from 0 to $i(s-r)$ and denotes which rectangle in the sum we're looking at. After rescaling, the heights are $\frac{1}{r+j/i}$. While j indicates how far to the right of $i \cdot r$ any given original rectangle lies, j/i indicates how far to the right of r any given rescaled rectangle lies. So think about the fact that the rectangle located at $r + j/i$ has a height of $\frac{1}{r+j/i}$. If we now just think of "located at" as "with $t =$" and "height of the rectangle" as "value of $f(t)$", it becomes clear that we are just integrating $f(t) = 1/t$ over the interval (r, s). We therefore have

$$\lim_{n\to\infty} P_q(n) = \lim_{n\to\infty} \sum_{t=\lceil q\cdot n\rceil}^{n} \frac{1}{t} = \lim_{i\to\infty} \sum_{t=i\cdot r}^{i\cdot s} \frac{1}{t} = \int_r^s \frac{1}{t}\, dt = \ln(s/r) = \ln(1/q)$$

Let's bring this back to the riddle. What we've just derived is the probability that any given card placement generates a loop that contains at least $q \cdot n$ cards for any $q : {}^1\!/_2 < q \leq 1$. And if the warden only allows each prisoner to open less than, but arbitrarily close to $q \cdot n$ boxes, then this is also the probability of failure. So the probability of success is simply $1 - \ln(1/q)$. And if the warden allows each prisoner to open exactly one-half of the boxes, then the chance of success is

$$\lim_{q\to {}^1\!/_2{}^+} (1 - \ln(1/q)) = 1 - \ln 2 \approx 0.307$$