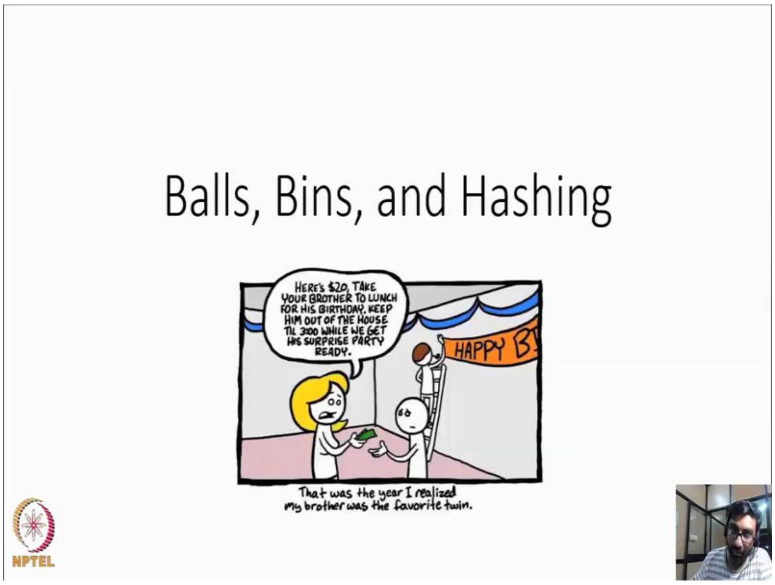
Algorithms for Big Data
Prof. John Ebenezer Augustine
Department of Computer Science and Engineering
Indian Institute of Technology, Madras

Lecture – 24
Balls, Bins, and Hashing

(Refer Slide Time: 00:14)



Balls, Bins, and Hashing

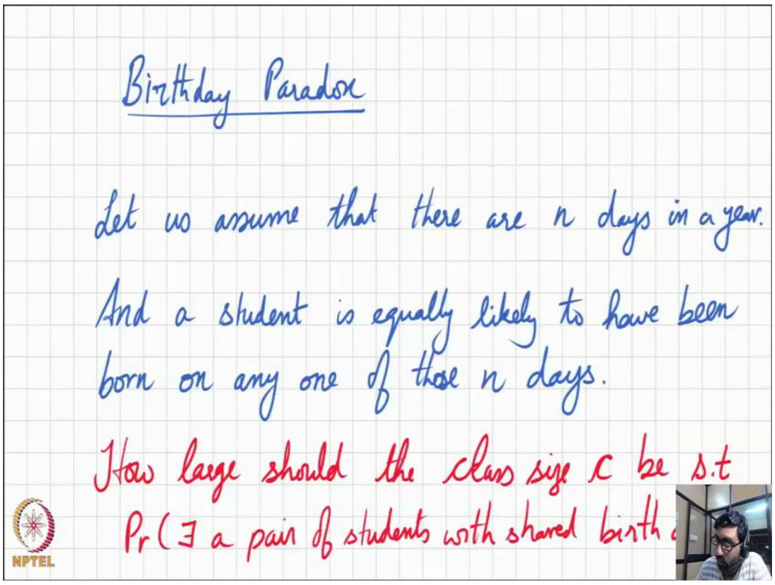
HERE'S $20, TAKE YOUR BROTHER TO LUNCH FOR HIS BIRTHDAY. KEEP HIM OUT OF THE HOUSE TIL 3:00 WHILE WE GET HIS SURPRISE PARTY READY.
HAPPY B
That was the year I realized my brother was the favorite twin.

Today’s lecture is about balls, bins and hashing. But first let us start with the funny paradox called the birthday paradox. So, the birthday paradox is typically concerned with a classroom full of students, there is some number of students. The question is what is the probability that there will be a pair of student with a same birthday?

Another way of another prospective one this birthday paradox is how large the class be for you expect to see a shared birthday, and at first sight you might think that you will need quite a few people in the class say about the say at least something like 100 students in the class because there are 365 days, but quite surprisingly somewhat at least you can except to see of shared birthday around at 23 students mark. So, when you get 23 students in the class you would likely to see a shared birthday, why is this so?

Let us think a little bit about that and in course of doing that we will develop an understanding of the balls and bins model, which should be helpful for us to visualize, what is happening in the birthday paradox? So, without further adieu, let us talk about the birthday paradox.

(Refer Slide Time: 01:48)



Birthday Paradox

Let us assume that there are n days in a year.

And a student is equally likely to have been born on any one of those n days.

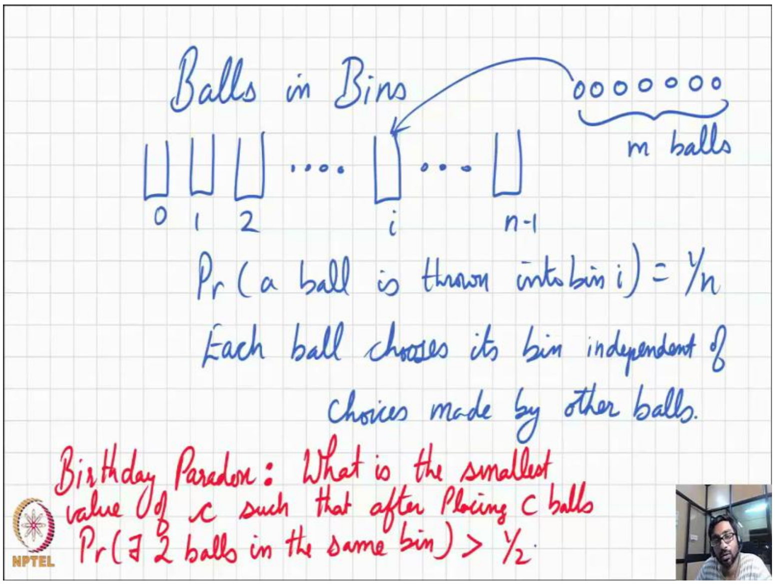
How large should the class size c be s.t

$\Pr(\exists$ a pair of students with shared birth

Let us begin with the assumption that there are n days in a year. So, this is more general than the 365 days that we are rather used to, but let us keep these general in order to help us understand how the flavor of balls and bins and now, let us assume also that a student is equally likely to have been born on any one of those n days. So, we are going to work with these assumptions.

So, the question that we are interested in is how large the class size be and let us denote the class size by c, how large should c be such that the probability that there exist a pair of students in that class of c students, who share a birthday and we want that probability at least a half and so let us see how this birthday paradox question can be analyzed using the balls and bins model.

(Refer Slide Time: 03:01)



Balls in Bins

Pr(a ball is thrown into bin i) = $1/n$

Each ball chooses its bin independent of choices made by other balls.

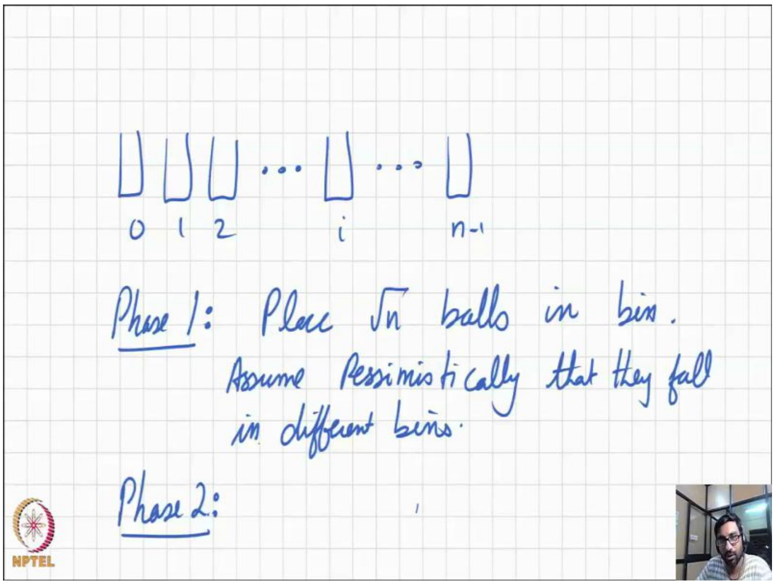
Birthday Paradox: What is the smallest value of c such that after placing c balls $Pr(\exists$ 2 balls in the same bin$) > 1/2$.

In this model, we have n bins. So, we identify these bins from 0 to n minus 1 and we have m ball and we assume that each of these balls sequentially thrown into one of the bins and the choice of bins is equally likely to be any one of the n bins and moreover the balls are thrown into bins chosen independently of each other.

Let us now try to rephrase the birthday paradox in the language of balls and bins. We are interested in the smallest number of balls that need to be thrown before the probability of a certain event becomes more than a half and what is the event? It is the event that 2 balls out of those c balls shared the same bin and this is exactly the birthday paradox except, we are now phrased it in the context of balls and bins.

Let us see how this question can be analyzed? On purpose we are going to be a little bit sloppy with our analysis mainly because we want to ensure that the intuition behind this paradox is brought out in a clear manner. So, the emphasis on the intuition, but you will notice if you look into the book there are many different ways in which this particular scenario can be analyzed. In fact, there are better ways more cleaner ways to analyze it, but we are going to choose this particular technique to analyze it only because it brings out the essence of the idea clear intuition behind it in the best possible way.

(Refer Slide Time: 04:55)



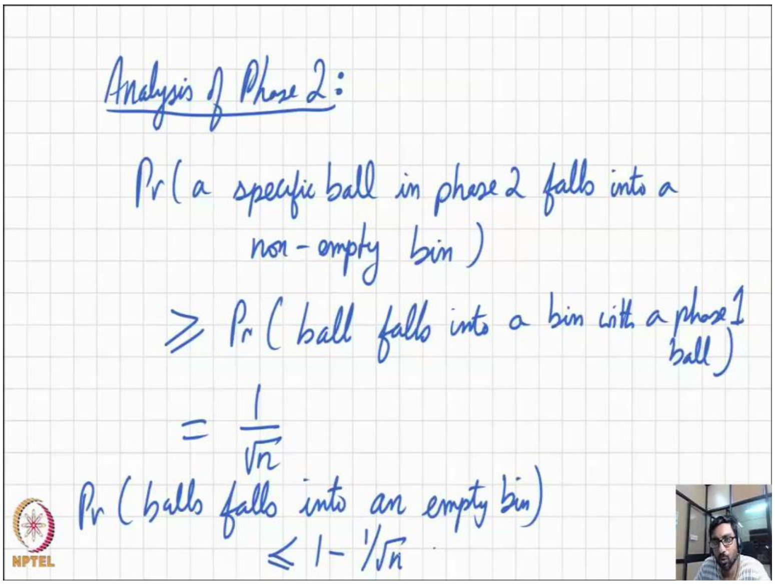
Phase 1: Place $\sqrt{n}$ balls in bin.

Assume Pessimistically that they fall in different bins.

Phase 2:

So, we have n bins, we are going to throw balls into 2 phases. In the first phase, which is going to throw square root of n balls and then there may be collations. So, they may be cases where 2 balls fallen in the same bin, but really we are concerned about getting a good lower bound on the probability that pair of balls will fall into the same bin. So, in a pessimistic way we going to assume that in the first phase there are no collations that these balls are fall into different bins. In phase 2, we will place the second square root of n balls, but now will have to be a bit more careful.

(Refer Slide Time: 05:48)



Analysis of Phase 2:

Pr(a specific ball in phase 2 falls into a non-empty bin)

$\geq$ Pr(ball falls into a bin with a phase 1 ball)

$= \frac{1}{\sqrt{n}}$

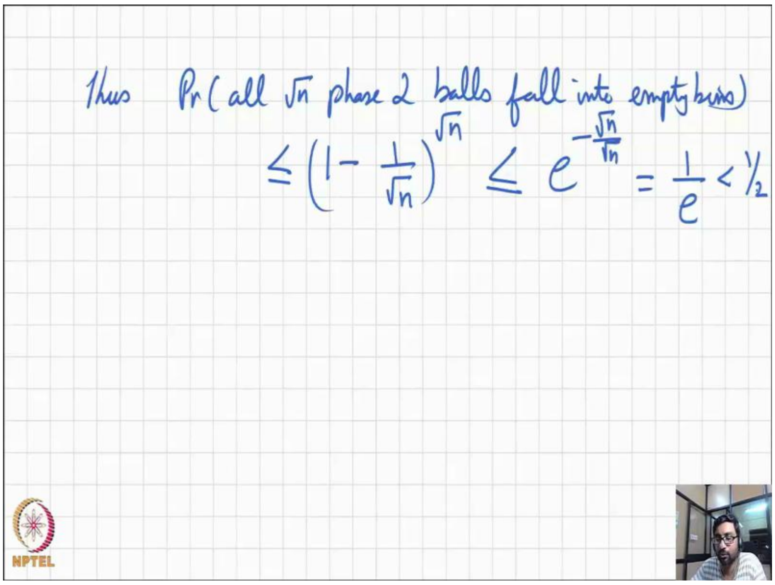
Pr(balls falls into an empty bin)

$\leq 1 - 1/\sqrt{n}$.

So, our entire focus is going to be on phase 2 and let us ask the question, what is the probability that a specific ball in phase 2 falls into a non-empty bin? Well, this probability is more than the probability that the ball falls into a bin with the phase 1 ball in it, why because there are going to be fewer bins with phase 1 balls than the total number of balls because this could be any ball in phase 2. So, the phase 2 could have brought in a few extra balls before our specific balls come into the picture.

So, we are going to ignore all of those balls. We are just going to bound this probability by the probability that this specific ball falls into a bin with phase 1 ball in it and this we know is equal to 1 over square root of n, which implies that the probability that the ball falls into an empty bin is at most 1 minus 1 over square root of n.

(Refer Slide Time: 07:14)



Thus $\Pr(\text{all } \sqrt{n} \text{ phase 2 balls fall into empty bins})$

$$\leq \left(1 - \frac{1}{\sqrt{n}}\right)^{\sqrt{n}} \leq e^{-\frac{\sqrt{n}}{\sqrt{n}}} = \frac{1}{e} < 1/2$$

Thus the probability that all the phase 2 ball fall in empty bins can be upper bounded by a 1 minus 1 over square root of n, the whole raise to the square root of n and of course, this is the at most e to the minus square root of n square root n and that is of course, 1 over e, which is certainly less than a half. So, clearly at most two times square root of n balls are required before the probability that we see a shared a bin for a pair of those balls exceeds half.

Of course, the birthday paradox is very well studied notions and as I mentioned there are several title analysis for it, but this balls and bins way of looking at the birthday paradox is particularly insightful. It is very easy to see, how the phase 2 balls operate in the presence of phase 1 ball, we are making a few pessimistic assumptions that kind of make it clear as to how the phase 2 balls behave and this square root of n the synthetic limit at least kind of comes out very cleanly, and also the as you might have noticed that the balls into bins model is pretty elementary, but is actually quite powerful in the number of ways in which it can be applied.

In this case, we have seen how the birthday paradox can be viewed as a balls and bins model, but of course, there are many other interesting problems especially algorithm problems that we are going be view as a instances of the balls and bins model, for example, we can view the coupon collectors problem as a variant of balls and bins. Hashing, which we are going to study now will be viewed as an instance of balls and bins. So, we are going to be able to reduce in a variety of ways. It really helps us to have a good understanding of how to analyze balls and bins questions.