THE BIRTHDAY PARADOX

Presentation by

Ankit Pradhan
16CS01014
IIT Bhubaneswar

The Problem

- Given a set of n randomly chosen people, what is the probability that a pair of these people have the same birthday?
- From ***Pigeonhole principle***, if n is 366 then the probability is 1 (as there are only 365 possible days in a normal year, excluding February 29 of leap year).
- Surprisingly, very high probabilities are obtained for even smaller values of n.

Analysis

- Let the probability that any two birthdays coincide be $p(n)$ and the probability that no two birthdays coincide be $\tilde{p}(n)$.

- $\tilde{p}(n) = 0$ for $n > 365$ and for $n \leq 365$,

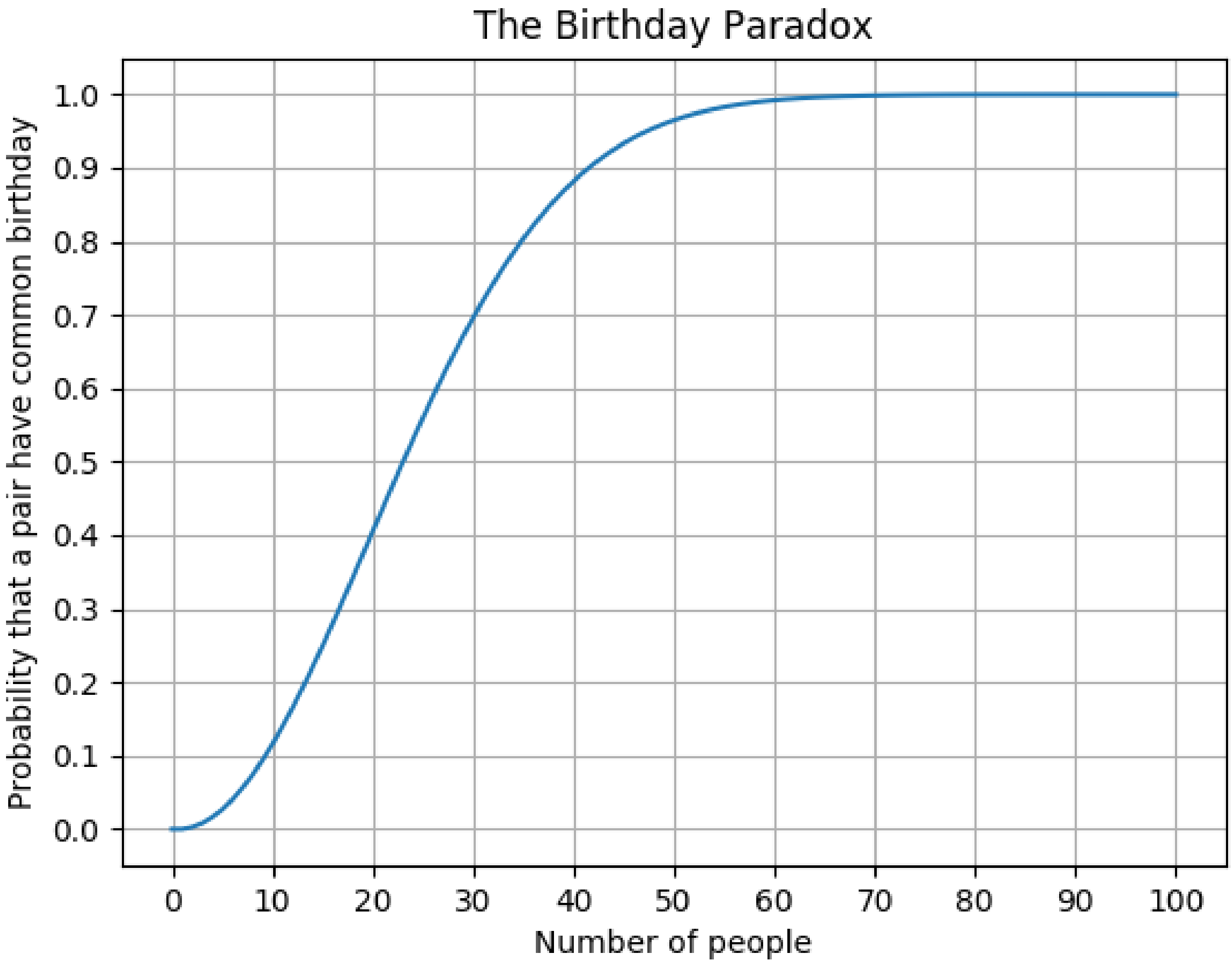
$$\tilde{p}(n) = \frac{365 \cdot 364 \cdot\cdots\cdot (365 - n+1)}{365^n} = \left(1 - \frac{1}{365}\right)\left(1 - \frac{2}{365}\right)\ldots\left(1 - \frac{n-1}{365}\right)$$

- Approximating $e^x \approx 1 + x$, for $x = -\frac{k}{365}$ gives $e^{-\frac{k}{365}} \approx 1 - \frac{k}{365}$

- $\tilde{p}(n) \approx e^{-\frac{1}{365}} \cdot e^{-\frac{2}{365}} \cdot\cdots\cdot e^{-\frac{n-1}{365}} \approx e^{-\frac{n(n-1)}{730}}$

- $\boldsymbol{p(n) = 1 - \tilde{p}(n) \approx 1 - e^{-\frac{n(n-1)}{730}}}$

- For $n = 23$, $p(n) \approx 0.5$, hence only 23 people are necessary to achieve 50% probability of two persons having same birthday.


The Birthday Paradox
Probability that a pair have common birthday
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
10
20
30
40
50
60
70
80
90
100
Number of people

Generalized Formulation

- The problem can be generalized as :

Given n random integers drawn from a discrete uniform distribution with range $[1, d]$, what is the probability $p(n, d)$ that at least two numbers are the same?

- It is easy to see that the required probability is

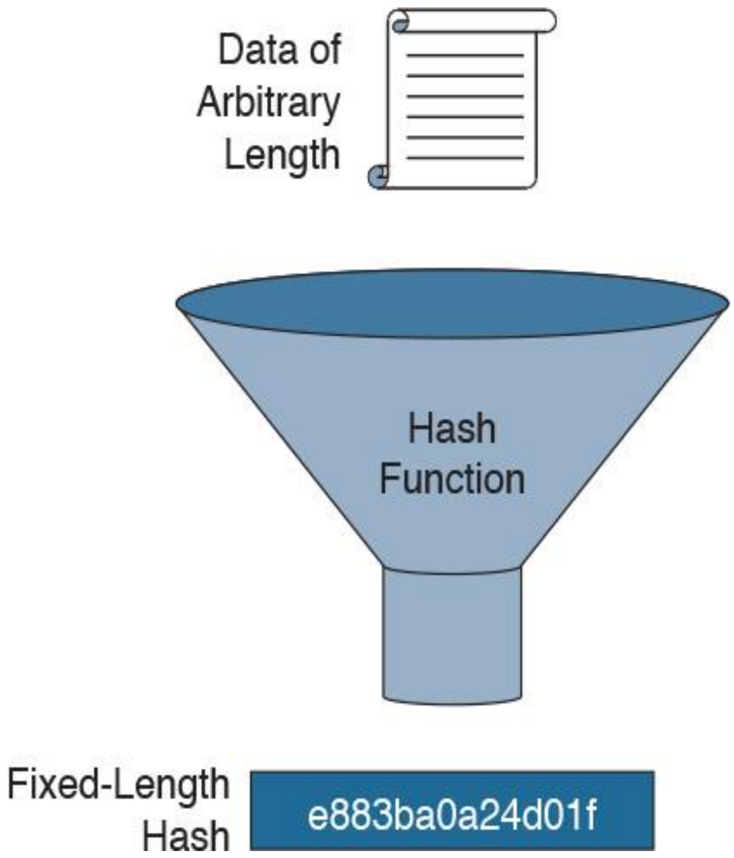
$$p(n,d) = \begin{cases} 1 - \prod_{k=1}^{n-1}\left(1 - \frac{k}{d}\right) & n \leq d \\ 1 & n > d \end{cases}$$

$$\approx 1 - \left(\frac{d - 1}{d}\right)^{\frac{n(n-1)}{2}}$$

$$\approx 1 - e^{-\frac{n(n-1)}{2d}}$$

Hash Functions

- A hash function is any function that maps data of arbitrary size to data of fixed length.
- Since the input size is variable, more than one input message can have the same hash digest.
- Cryptographic Hash Functions are designed to be non-invertible and provide collision resistance.



Collision Resistance and Birthday Paradox

- A Cryptographic Hash Function H is collision resistant if it is hard to find two distinct inputs x, y $(x \neq y)$ such that $H(x) = H(y)$.

- From the generalized birthday paradox,

$$p(n, d) \approx 1 - e^{-\frac{n(n-1)}{2d}} \approx 1 - e^{-\frac{n^2}{2d}}$$

- If $n(p, d)$ denotes the number of random integers drawn from $[1, d]$ to obtain a probability p that at least two numbers are same, then

$$n(p, d) \approx \sqrt{2d \cdot ln\left(\frac{1}{1-p}\right)}$$

- For $p = 1/2$, $n \approx \sqrt{2d \cdot \ln(2)} \approx 1.18 \times d^{1/2}$.

Birthday Attack

- Let $H : M \rightarrow \{0,1\}^n$ be a hash function with M being the message space and n being the size of hash digest.

- Algorithm to find a collision in time $O(2^{n/2})$ hashes
  1. Choose $2^{n/2}$ distinct random messages in M: $m_1, \ldots, m_{2^{n/2}}$
  2. For $i = 1, \ldots, 2^{n/2}$ compute $t_i = H(m_i)$
  3. Look for a collision ($t_i = t_j$) for $i \neq j$. If not found, goto 1.

- This comes from the fact that given n distinct numbers $r_1, \ldots, r_n$ from a sample of size d, probability $P[\exists\, i \neq j : r_i = r_j] \geq \frac{1}{2}$ holds if $n \geq 1.18 \times d^{1/2}$.

REFERENCES

- https://en.wikipedia.org/wiki/Birthday_problem
- https://en.wikipedia.org/wiki/Collision_resistance
- https://en.wikipedia.org/wiki/Cryptographic_hash_function
- https://www.coursera.org/learn/crypto/lecture/pyR4I/generic-birthday-attack

THANK YOU

Ankit Pradhan