Cryptography: Birthday Paradox

Yiheng Lin, Zhihao Jiang

Introduction and Outline

In this project, we study the Birthday Paradox, in which we uniformly randomly draw n elements with replacement from a set of N elements. The primary goal is to prove a bound of n to promise a collision (some element is drawn for at least 2 times) with some constant probability p. The result is n should be at least $\theta(N^{\frac{1}{2}})$ and we proved this bound cannot be improved. We study this in the first part.

In the second part, we prove that if the probability of drawing each one of the N elements is not uniformly $\frac{1}{N}$, the probability that a collision occurs is greater or equal to the probability under the uniform case. This provide a plausible reason on why we should use hash functions that have uniform probability and how we can attack them.

Then we study the bound of n to promise a d-time collision (some element is drawn for at least d times) with some probability p. The result is n should be at least $\theta(N^{\frac{d-1}{d}})$ and we proved this bound cannot be improved.

1

1.1

Theorem 1.1. Let $S = \{1, 2, \cdots, N\}$. For n times, uniformly randomly draw one element from set S with replacement. Let x_t be the element we draw at time t. Then $\forall p > 0$, there exists a constant C_1 such that when $n \geq C_1\sqrt{N}$, we have

$$Pr[\exists 1 \leq i, j \leq n, i \neq j \text{ such that } x_i = x_j] > p.$$

Proof. Let X denote the event that $\exists i, j \leq t, i \neq j$ such that $x_i = x_j$, then we have

$$\begin{aligned} Pr[\bar{X}] &= \frac{N(N-1)\cdots(N-n+1)}{N^n} \\ &= \prod_{i=1}^{n-1}(1-\frac{i}{N}) \\ &\leq \prod_{i=1}^{n-1} exp(-\frac{i}{N}) \\ &= exp(-\frac{n(n-1)}{2N}). \end{aligned} \tag{1}$$

Let $C_1 = \sqrt{-2ln(1-p)} + 1$. Then when $n \geq C_1\sqrt{N}$, we have

$$n(n-1) > (1 + \sqrt{-2ln(1-p) \cdot N})(\sqrt{-2ln(1-p) \cdot N}) > -2ln(1-p) \cdot N. \tag{2}$$

Which is equivalent to $-\frac{n(n-1)}{2N} < ln(1-p)$. Thus use (1) we have

$$Pr[\bar{X}] \leq exp(-\frac{n(n-1)}{2N}) < 1 - p.$$

So we have

$$Pr[X] > p$$

□

1.2

Lemma 1.1. For positive integer $n < N$, we have

$$\sum_{i=1}^{n-1} ln(1-\frac{i}{N}) > -\frac{n^2}{N}.$$

Proof. Notice that $\forall x \in [1-\frac{i+1}{N}, 1-\frac{i}{N}]$ $(0 \leq i \leq n)$, we have $ln(x) < ln(1-\frac{i}{N})$. Thus

$$\frac{1}{N} ln(1-\frac{i}{N}) \geq \int_{1-\frac{i+1}{N}}^{1-\frac{i}{N}} ln(x)dx.$$

Thus we have

$$\begin{aligned} \frac{1}{N}\sum_{i=1}^{n-1} ln(1-\frac{i}{N}) &\geq \int_{1-\frac{n}{N}}^{1} ln(x)dx \\ &= (xln(x)-x)|_{1-\frac{n}{N}}^{1} \\ &= -\frac{n}{N} - (1-\frac{n}{N})ln(1-\frac{n}{N}) \\ &> -\frac{n}{N} - (1-\frac{n}{N})(-\frac{n}{N}) \\ &= -\frac{n^2}{N^2}. \end{aligned} \tag{3}$$

Thus

$$\sum_{i=1}^{n-1} ln(1-\frac{i}{N}) > -\frac{n^2}{N}.$$

□

Theorem 1.2. Let $S = \{1, 2, \cdots, N\}$. For n times, uniformly randomly draw one element from set S with replacement. Let x_t be the element we draw at time t. Then $\forall p > 0$, there exists a constant C_2 such that when $n \leq C_2\sqrt{N}$, we have

$$Pr[\exists 1 \leq i, j \leq n, i \neq j \text{ such that } x_i = x_j] < p.$$

Proof. Let X denote the event that $\exists i, j \leq t, i \neq j$ such that $x_i = x_j$.

Use Lemma 1.1, we have

$$\begin{aligned} Pr[\bar{X}] &= \frac{N(N-1)\cdots(N-n+1)}{N^n} \\ &= \prod_{i=1}^{n-1}(1-\frac{i}{N}) \\ &= exp(\sum_{i=1}^{n-1} ln(1-\frac{i}{N})) \\ &> exp(-\frac{n^2}{N}). \end{aligned} \tag{4}$$

Let $C_2 = \sqrt{-ln(1-p)}$. Then when $n \leq C_2\sqrt{N}$, we have

$$exp(-\frac{n^2}{N}) \geq 1-p.$$

So $Pr[\bar{X}] > 1 - p$, thus

$$Pr[X] < p.$$

□

2

Theorem 2.1. Let $S = \{1, 2, \cdots, N\}$. Let $D_1 : S \to R^+ \cup \{0\}$ be a discrete probability distribution over S. For n times, randomly draw one element from set S according to distribution D_1 with replacement . Let x_t be the element we draw at time t. Let D_0 be the uniform distribution over S, which satisfies $\forall i \in S, D_0(i) = \frac{1}{N}$. Then we have

$$Pr_{D_1^n}[\exists 1 \leq i, j \leq n, i \neq j \text{ such that } x_i = x_j] \geq Pr_{D_0^n}[\exists 1 \leq i, j \leq n, i \neq j \text{ such that } x_i = x_j].$$

Proof. Let X denote the event that $\exists i, j \leq t, i \neq j$ such that $x_i = x_j$. Let X_m denote the event that $\exists 1 \leq i, j \leq n, i \neq j$ such that $x_i = x_j = m$.

First, to change D_1 to D_0, we can apply the following algorithm:

```
$t := 1$
While $D_t \neq D_0$:
    find $i, j \in S$ such that $D_t[i] < \frac{1}{N} < D_t[j]$
    let $D_{t+1}[j] := D_t[i] + D_t[j] - \frac{1}{N}, D_{t+1}[i] := \frac{1}{N}, \forall k \neq i, j, D_{t+1}[k] := D_t[k]$
    $t++$
End While
```

Since the number of $\frac{1}{N}$ in D increases at each iteration, this algorithm will terminate in N steps.

We only need to prove that

$$\forall t, Pr_{D_t^n}[X] \geq Pr_{D_{t+1}^n}[x].$$

Without losing generality, suppose when generate D_{t+1} from D_t, we choose $i = 1, j = 2$.

Let Y be the number of times that the element we draw is in $\{1, 2\}$.

$$\begin{aligned} Pr_{D^n}[X] &= Pr_{D^n}[\bigcup_{k=3}^{N} X_k] + (1 - Pr_{D^n}[\bigcup_{k=3}^{N} X_k])Pr_{D^n}(X_1 \cup X_2 | \bigcap_{k=3}^{N} \bar{X}_k) \\ &= Pr_{D^n}[\bigcup_{k=3}^{N} X_k] + (1 - Pr_{D^n}[\bigcup_{k=3}^{N} X_k]) \sum_{i=0}^{\infty} Pr_{D^n}(Y = i | \bigcap_{k=3}^{N} \bar{X}_k) Pr_{D^n}(X_1 \cup X_2 | Y = i). \end{aligned} \tag{5}$$

The last equation holds because $\forall i, Pr_{D^n}(X_1 \cup X_2 | Y = i) = Pr_{D^n}(X_1 \cup X_2 | Y = i, \bigcap_{k=3}^{N} \bar{X}_k)$.

Notice that

$$\begin{aligned} & Pr_{D_t^n}(X_1 \cup X_2 | Y = 2) - Pr_{D_{t+1}^n}(X_1 \cup X_2 | Y = 2) \\ &= \frac{1}{(D_t[1] + D_t[2])^2}(D_t[1]^2 + D_t[2]^2 - (\frac{1}{N})^2 - (D_t[1] + D_t[2] - \frac{1}{N})^2) \\ &= -\frac{2}{(D_t[1] + D_t[2])^2}(D_t[1] - \frac{1}{N})(D_t[2] - \frac{1}{N}) \\ &> 0. \end{aligned} \tag{6}$$

And for any distribution D over S we have

$$Pr_{D^n}(X_1 \cup X_2 | Y = i) = \begin{cases} 0 & i = 0, 1 \\ 1 & i \geq 3 \end{cases}. \tag{7}$$

Thus

$$\forall i, Pr_{D_t^n}(X_1 \cup X_2 | Y = i) \geq Pr_{D_{t+1}^n}(X_1 \cup X_2 | Y = i).$$

Since we only adjust $D_t[1], D_t[2]$,

$$Pr_{D_t^n}[\bigcup_{k=3}^{N} X_k] = Pr_{D_{t+1}^n}[\bigcup_{k=3}^{N} X_k]$$

$$\forall i, Pr_{D_t^n}(Y = i | \bigcap_{k=3}^{N} \bar{X}_k) = Pr_{D_{t+1}^n}(Y = i | \bigcap_{k=3}^{N} \bar{X}_k).$$

So consider equation (5), (6) and we get

$$Pr_{D_t^n}[X] \geq Pr_{D_{t+1}^n}[x].$$

□

3

3.1

Theorem 1. *Let $S = \{1, 2, \cdots, N\}$. For n times, uniformly randomly draw one element from set S with replacement. Let x_t be the element we draw at time t. Then for all integer $d \geq 2$ and for all $p > 0$, there exists a constant C_1 such that when $n \geq C_1 N^{\frac{d-1}{d}}$, we have*

$$Pr[X] > p,$$

where X denotes the event $\exists 1 \leq i_1 < i_2 < \cdots < i_d \leq n$, such that $x_{i_1} = x_{i_2} = \cdots = x_{i_d}$.

Proof. We prove this theorem by induction. This theorem is right when $d = 2$ which is proved before.

The choice of C_1 is dependent of p and d, we denote the constant as $C_1(p, d)$ in this proof.

Now assume the theorem is right when $d = k - 1$, and we prove the theorem is right when $d = k$.

By induction, $\forall p$, we can find a constant C_1 (to make it more convenient, we do not use notation $C_1(\frac{p+1}{2}, d)$ here, but make sure constants should be independent with N) such that

$$Pr[X] > \frac{1+p}{2}$$

for all N. Let C_2 be another constant such that $\frac{1+p}{2} \cdot (1 - (\frac{1}{e})^{C_2}) > p$ and $(1 - exp(-\frac{C_1+C_2}{4})) > p$. We divide the whole drawing process into two steps:

1. First, draw $M_1 = C_1 \cdot N^{\frac{d-2}{d}} + C_2 \cdot N^{\frac{d-1}{d}}$ elements from S. Let set A be the set of all the elements that has been drawn for at least once.

2. Second, draw $M_2 = 2(C_1 + C_2) \cdot N^{\frac{d-1}{d}}$ elements from S. Let Y be the number of times that an element is drawn from set A.

Now we consider 2 possible cases of the size of A.

3.1.1 First Case

If $|A| \leq N^{\frac{d-1}{d}}$:

By assumption, after drawing for $C_1 \cdot N^{\frac{d-2}{d}}$ times, let event E_1 be that there exists an element c_0 in A that has been drawn for at least $d - 1$ times, by assumption, given that $|A| \leq N^{\frac{d-1}{d}}$, we have $Pr(E_1) > \frac{p+1}{2}$.

Notice that given a fixed element c in A, the probability that $|A|$ random draws draw c for 0 times is $(1 - \frac{1}{|A|})^{|A|} < \frac{1}{e}$. So let the event E_2 be that the element c_0 has been drawn for at least once in the last $C_2 \cdot N^{\frac{d-1}{d}}$ draws. Then we have the conditional probability $P(E_2|E_1) > (1 - (\frac{1}{e})^{C_2})$.

If E_1 and E_2 both happens, then c_0 must be drawn for at least $(d-1)+1 = d$ times. And the joint probability is

$$Pr(E_1, E_2) = Pr(E_1) \cdot Pr(E_2|E_1) > \frac{p+1}{2} \cdot (1 - (\frac{1}{e})^{C_2}) > p. \tag{8}$$

Thus we have proved that

$$Pr[X||A| \leq N^{\frac{d-1}{d}}] > p.$$

just after the first step.

3.1.2 Second Case

Else, we have $(C_1 + C_2)N^{\frac{d-1}{d}} > |A| > N^{\frac{d-1}{d}}$:

Now we try to bound the probability that $Y < (|A|)^{\frac{d-2}{d-1}}$ in step 2.

Let X_i be the indicater random variable of whether the i th draw in step 2 draws an element in A. In other words,

$$X_i = \begin{cases} 1 & \text{if the i th draw draws an element from A} \\ 0 & \text{otherwise} \end{cases}. \tag{9}$$

Then we have $Y = \sum_{i=1}^{M_2} X_i$. And X_is are i.i.d. Bornoulli random variables. So use Chernoff Bound

$$Pr(X < (1-\delta)\mu) \leq e^{-\frac{\delta^2 \mu}{2}}$$

where $0 \leq \delta \leq 1$, X is the sum of the random variables, μ is the expected value of the sum. We have

$$Pr[Y \leq \frac{1}{2}E[Y]] \leq e^{-\frac{E[Y]}{8}}. \tag{10}$$

Here

$$E[Y] = \frac{|A|}{N} \cdot 2(C_1 + C_2) \cdot N^{\frac{d-1}{d}} \geq 2 \cdot (|A|)^{\frac{d-2}{d-1}}.$$

And

$$E[Y] = \frac{|A|}{N} \cdot 2(C_1 + C_2) \cdot N^{\frac{d-1}{d}} \geq 2 \cdot (C_1 + C_2) \cdot N^{\frac{d-2}{d}}.$$

Thus we have

$$Pr[Y > (|A|)^{\frac{d-2}{d-1}}] \geq 1 - exp(-\frac{(C_1 + C_2) \cdot N^{\frac{d-2}{d}}}{4}) \geq 1 - exp(-\frac{C_1 + C_2}{4}).$$

Let the event E_3 be that there exists an elemnet c_1 in A such that c_1 has been drawn for at least $d-1$ times in step 2. By induction, we know in step 2, the conditional probability

$$Pr[E_3|Y > (|A|)^{\frac{d-2}{d-1}}] > \frac{1+p}{2}.$$

Thus we have

$$Pr[E_3] = Pr[E_3|Y > (|A|)^{\frac{d-2}{d-1}}] \cdot Pr[Y > (|A|)^{\frac{d-2}{d-1}}] > \frac{p+1}{2} \cdot (1 - exp(-\frac{C_1 + C_2}{4})) > p. \tag{11}$$

Since the event E_3 gaurantees that c_1 has been drawn for at least d times (at least (d-1) in step 2, and at least 1 in step 1), we proved that

$$Pr[X||A| > N^{\frac{d-1}{d}}] > p.$$

Combining Subsection 3.1.1 and Subsection 3.1.2, we get $Pr[X] > p$ for $d = k$. Thus we have finished the proof by induction. □

3.2

Theorem 2. *Let $S = \{1, 2, \cdots, N\}$. For n times, uniformly randomly draw one element from set S with replacement. Let x_t be the element we draw at time t. Then for all integer $d \geq 2$ and for all $p > 0$, there exists a constant C_2 such that when $n \leq C_2 N^{\frac{d-1}{d}}$, we have*

$$Pr[X] < p,$$

where X denotes the event $\exists 1 \leq i_1 < i_2 < \cdots < i_d \leq n$, such that $x_{i_1} = x_{i_2} = \cdots = x_{i_d}$.

Proof. Let $C_2 = \sqrt[d]{p}$. We have

$$\begin{aligned} Pr[X] &\leq \sum_{i_1=1}^{C_2 n} \sum_{i_2=i_1+1}^{C_2 n} \cdots \sum_{i_d=i_{d-1}+1}^{C_2 n} Pr[x_{i_1} = x_{i_2} = \cdots = x_{i_d}] \\ &= \sum_{i_1=1}^{C_2 n} \sum_{i_2=i_1+1}^{C_2 n} \cdots \sum_{i_d=i_{d-1}+1}^{C_2 n} \frac{1}{N^{d-1}} \\ &< \frac{C_2^d n^d}{N^{d-1}} \\ &= p. \end{aligned}$$

□

4 Future Research

A shortage of our work is that we did not impose tight bound on the constant C_1 and C_2 as a function of p and d. In another word, the function $C_1(p, d)$ could be exponential with respect to parameters p and d. Professor Wu mentioned that another group (Chi Han et al.) has considered this issue carefully by deriving a recursive relationship on function $p(n, d)$ (express the probability of d-time collision as a function of n and d). Their result is promising and we will discuss with them in the future.

Another thing Professor Wu asked us to consider is study when non-uniform probability will occur in common hash functions.

Actually, this may happen when the input data itself has some special properties. For example, we know the multi-way cache in computer uses a part of digits in the physical addresses as index. The designer choose the middle part of the address (between tag and page offset) because they think this part usually has good randomness. But if a program accidently access addresses which have the same middle part, the randomness will be broken and the miss rate increases dramatically.

And the differential attack can be viewed as using non-uniform probability distribution in hash function. Over uniform distribution of difference of plaintexts, the distribution of ciphertexts (or hash values) is not uniform. Some difference of ciphertexts (or hash values) may happen more frequently than others. The differential attack can utilize this property.

Reference

- Concrete Mathematics: A Foundation for Computer Science, Ronald Graham, Donald Knuth, and Oren Patashnik, 1989

- A First Course in Proability, Sheldon M. Ros

- Understanding Cryptography, Christof Paar · Jan Pelzl