Cryptography Project
Birthday Paradox

Lu Rui, Han Chi

January 11, 2019

Contents

1 Rigorous Proof for Birthday Paradox Bound **2**
1.1 Recursive BP Theorem 2
1.2 Upper and Lower Bound for $Pr[C(n,q,s)]$ 3
1.3 Explanation for $q = \Theta(n^{(s-1)/s})$ 4

2 Special for $s = 2$ **5**

3 Some Interesting Things **5**

4 Situation for $q = \Theta\left(n^{\frac{s-1}{s}}\right)$ and the constant **6**
4.1 Bounds for $q = \Theta\left(n^{(s-1)/s}\right)$ 6
4.2 Tight Bound of $s-$collision 7

5 Near-Hits **8**
5.1 Original results 8
5.2 Generalization to continuous space 9

6 Non-uniform Distribution **10**
6.1 Main theorem and conclusions 10
6.2 Analysis: How to avoid collisions 12
6.3 Proof of the theorem 12
6.4 Generalization: m-th collisions and s-way collision 14
6.4.1 Distribution for further collisions 14
6.4.2 m-th collision 15
6.4.3 s'-way collision 17

7 Future work **18**

8 Appendix **18**
8.1 Proof for Trivial lemma 18

1 Rigorous Proof for Birthday Paradox Bound

In this section, we are going to analyze the asypmtotic bound for the expected number of sample to get a collision for a fully random hash function.

Suppose that we have a hash function $H(x) : \mathcal{D} \mapsto \mathcal{R}$, which maps from domain $\mathcal{D}$ to range $\mathcal{R}, |\mathcal{R}| = n$. We use $Pr[C(n, q, s)]$ to denote the probability that one can get a s-way hash collision within a discrete domain of size n by generating q different variables. We will first prove several lemmas:

1.1 Recursive BP Theorem

This theorem is vital for us to fully analyze the exactly probability exactly.

Theorem 0

$$Pr[C(n,q,s)] = \frac{1}{n^{s-1}} \sum_{i=s}^{q} \binom{i-1}{s-1} \left(1 - \frac{1}{n}\right)^{i-s} (1 - Pr[C(n-1, i-s, s)])$$

Proof. We consider this as throwing q balls randomly into n buckets. For each $s \leq i \leq q$, use $C(n, q, s, i)$ denote the event that the i^{th} ball causes the first s-collision. Then since all those events are exclusive, we can conclude from the union bound that

$$Pr[C(n,q,s)] = \sum_{i=s}^{q} Pr[C(n,q,s,i)]$$

To calculate the probabililty for each $Pr[C(n, q, s, i)]$, we enumerate the condition step by step

- Determine the bucket to contain the collision, say B, of which there are n possibilities.
- Apart from the s^{th} ball which is the last, we need to determine when do the $s-1$ balls among the $i-1$ balls get into B. This contributes $\binom{i-1}{s-1}$ variabilities, and requiring all those cillision balls to fall in B contributes probability $\frac{1}{n^s}$.
- Then for the rest of the balls, they are not allowed to get into B, of which each ball happen this with probability $1 - \frac{1}{n}$.
- For the rest $i-s$ balls which fallen into other $n-1$ boxes, the probability that they do not collide happens with probability $1 - Pr[C(n-i, i-s, s)]$.

Then, conclude all those steps by multiplication rule, we have the given recursive equation. □

1.2 Upper and Lower Bound for $Pr[C(n,q,s)]$

First state the bound as theorem 1

Theorem 1

$$Pr[C(n,q,s)] \leq \frac{1}{n^{s-1}} \binom{q}{s}$$

$$Pr[C(n,q,s)] \geq \frac{1}{n^{s-1}} \binom{q}{s} \left(1-\frac{1}{n}\right)^{q-s} \left[1 - \frac{1}{2(n-1)^{s-1}} \binom{q-s}{s}\right]$$

This theorem fully characterizes the probability of collision for a $s-$collision.

Proof. According to theorem 0,

$$\begin{aligned}
\Pr[C(n,q,s)] &= \frac{1}{n^{s-1}} \sum_{i=s}^{q} \left(\begin{array}{c} i-1 \\ s-1 \end{array} \right) \left(1-\frac{1}{n}\right)^{i-s} (1 - \Pr[C(n-1, i-s, s)]) \\
&\leq \frac{1}{n^{s-1}} \sum_{i=s}^{q} \left(\begin{array}{c} i-1 \\ s-1 \end{array} \right) \\
&= \frac{1}{n^{s-1}} \left(\begin{array}{c} q \\ s \end{array} \right)
\end{aligned}$$

The last equation is a common lemma in combinatorics.
Also on the another side we have

$$\begin{aligned}
Pr[C(n,q,s)] &= \frac{1}{n^{s-1}} \sum_{i=s}^{q} \binom{i-1}{s-1} \left(1-\frac{1}{n}\right)^{i-s} (1 - Pr[C(n-1, i-s, s)]) \\
&\geq \frac{1}{n^{s-1}} \left(1-\frac{1}{n}\right)^{q-s} \sum_{i=s}^{q} \binom{i-1}{s-1} (1 - Pr[C(n-1, i-s, s)]) \\
&= \frac{1}{n^{s-1}} \left(1-\frac{1}{n}\right)^{q-s} \left[\binom{q}{s} - \sum_{i=2s}^{q} \binom{i-1}{s-1} Pr[C(n-1, i-s, s)]) \right]
\end{aligned}$$

The last equation comes from $Pr[C(n-1, i-s, s)] = 0$ for $i - s < s$. Now

that we can plug in the upper bound above to reach the lower bound.

$$\begin{aligned} Pr[C(n,q,s)] &\geq \frac{1}{n^{s-1}}\left(1-\frac{1}{n}\right)^{q-s}\left[\binom{q}{s}-\sum_{i=2s}^{q}\binom{i-1}{s-1}\frac{1}{(n-1)^{s-1}}\binom{i-s}{s}\right] \\ &= \frac{1}{n^{s-1}}\left(1-\frac{1}{n}\right)^{q-s}\left[\binom{q}{s}-\frac{1}{(n-1)^{s-1}}\binom{2s-1}{s}\binom{q}{2s}\right] \\ &= \frac{1}{n^{s-1}}\left(1-\frac{1}{n}\right)^{q-s}\binom{q}{s}\left[1-\frac{1}{2(n-1)^{s-1}}\binom{q-s}{s}\right] \end{aligned}$$

Proof of the second line to the last can be found in Appendix section. □

From now on, we will denote

$$f(n)=\left(1-\frac{1}{n}\right)^{q-s} \text{ and } g(n)=\frac{1}{2(n-1)^{s-1}}\binom{q-s}{s}$$

Then we can restate theorem 1 in a more compact way as

$$f(n)(1-g(n))\frac{1}{n^{s-1}}\binom{q}{s} \leq Pr[C(n,q,s)] \leq \frac{1}{n^{s-1}}\binom{q}{s}$$

1.3 Explanation for $q=\Theta(n^{(s-1)/s})$

Before we dive into the deduction for the bounds above, first explain the intuition and what it tells us. Notice that the difference between two side is just a proportion term $f(n)(1-g(n))$, we wish to prove that with the growth of n, the left part can be arbitrarily close to 1. Hence this implies that to reach a collision, the probability $Pr[C(n,q,s)] := p(q)$ with respect to q is essentially a polynomial with degree s for term $\binom{q}{s}$. And approximately, to let the term $\frac{1}{n^{s-1}}\binom{q}{s}$ approach a constant non-negligible probability, q^s has to be about n^{s-1}, which means q should be about $\Theta(n^{(s-1)/s})$

Then, we will formally prove this result, namely we will show that for any constant $c<1$, if $q=n^{\epsilon}, \epsilon < \frac{s-1}{s}$, then there always exists a sufficiently large n_0 s.t. for $\forall n \geq n_0$

$$f(n)(1-g(n)) > c$$

Proof.

$$g(n) < \frac{1}{2s!}\frac{q^s}{n^{s-1}} = \frac{1}{2s!n^{s-1-s\epsilon}} \to 0$$

$$
\begin{aligned}
f(n) &= \left(1-\frac{1}{n}\right)^{q-s} > \left(1-\frac{1}{n}\right)^{q} \\
&= \left[\left(1-\frac{1}{n}\right)^{-n}\right]^{-\frac{q}{n}} \\
&= e_n^{-n^{\epsilon-1}}
\end{aligned}
$$

Here $e_n = \left(1-\frac{1}{n}\right)^{-n} \to e$. Since $\epsilon < \frac{s-1}{s} < 1$, hence $\epsilon - 1 < 0$ and we know $f(n) \to e^0 = 1$ □

Now, combine the statements above we know that, if $\epsilon < \frac{s-1}{s}$, then for any $q = O(n^\epsilon)$, the probability for collision will converge to exactly $Pr[C(n,q,s)] \sim \frac{1}{n^{s-1}}\binom{q}{s} \leq \frac{q^s}{s!n^{s-1}}$. And since $\epsilon < \frac{s-1}{s}$ we have $q^s = o(n^{s\epsilon}) = o(n^{s-1})$, so $Pr[C(n,q,s)] \to 0$ for $q = o(n^{(s-1)/s})$ when $n \to +\infty$. That indicates the infimum for the degree of drawing random times is $\frac{s-1}{s}$.

2 Special for $s = 2$

Denote event "a hash collision happens" as A. Then we have

$$
\begin{aligned}
Pr[A] &= 1 - Pr[\bar{A}] = 1 - \frac{n(n-1)(n-2)...(n-q+1)}{n^q} \\
&= 1 - \prod_{i=1}^{q-1}\left(1-\frac{i}{n}\right) \geq 1 - \prod_{i=1}^{q-1} e^{-i/n} \\
&= 1 - e^{-\frac{q(q-1)}{2n}} \geq 1 - e^{-\frac{(q-1)^2}{2n}}
\end{aligned}
$$

And set $1 - e^{-\frac{(q-1)^2}{2n}} = 0.5$, we have $q \approx 1.18\sqrt{n} + 1$ will yield a collision with probability over 0.5.

3 Some Interesting Things

Despite the tedious deduction above, how can we directly get the insight about the degree $\frac{s-1}{s}$ for a s-collision? The answer is actually quite simple.

Consider q random draws from a finite n set. There are $\binom{q}{s}$ number of s-tuple within such q random draws. For each tuple, the probability they collide is $\frac{n}{n^s} = \frac{1}{n^{s-1}}$. By the linearity of expectation, we know that the expected number for s-collision is $\binom{q}{s} \times \frac{1}{n^{s-1}}$. So approximately, to make such a collision happen, q has to create about $\frac{1}{1/n^{s-1}} = n^{s-1}$ number

of $s-$tuple, which means $\binom{q}{s} = n^{s-1}$. The left part can be viewed as a polynomial for q of degree s, of which the first coefficient is $\frac{1}{s!}$. Hence the expected number of draws to make a $s-$collision is about $\frac{q^s}{s!} \approx n^{s-1}$, which means $q \sim s! n^{(s-1)/s}$.

4 Situation for $q = \Theta\left(n^{\frac{s-1}{s}}\right)$ and the constant

Now that we know to make a collision, the degree of $\frac{s-1}{s}$ to n is necessary. This means to reach non-negligible collision probability, we need $\Omega\left(n^{\frac{s-1}{s}}\right)$ number of samples of get a collision. In this section, we will show $q = O(n^{\frac{s-1}{s}})$, and fully investigate the constant before it, namely we need

$$q \approx (s!)^{1/s} n^{(s-1)/s} + s - 1$$

This result shows that when s is large, merely sample $n^{(s-1)/s}$ times is not enough. By our previous argument, this could only make a collision with probability $\frac{1}{s!}$, which shrinks with s rapidly.

4.1 Bounds for $q = \Theta\left(n^{(s-1)/s}\right)$

Theorem 2

$$Pr[C(n,q,s)] \leq \frac{1}{n^{s-1}} \binom{q}{s} < \frac{\alpha^s}{s!} < \frac{1}{s!}$$

Denote $\alpha = \frac{q}{n^{(s-1)/s}}, \alpha' = \frac{q-s}{n^{(s-1)/s}}$, of which $0 < \alpha' < \alpha < 1$. Denote $e_n = \left(1 + \frac{1}{n}\right)^n$. If $2 \leq s \leq q$ then we have

$$Pr[C(n,q,s)] > \frac{\alpha'^s}{s!} - \left(\frac{\alpha'^{s+1} \ln(e_n)}{s! n^{1/s}} + \frac{\alpha^s \alpha'^s}{2(s!)^2}\right)$$

Theorem 2 gives a concrete and accurate bound for the probability of a $s-$way collision if we sample just for $\alpha n^{(s-1)/s}$ where $\alpha < 1$. Notice that when $n \to \infty$, $ln(e_n) \to 1$ and $\alpha' \to 1, \alpha \to 1$, the lower bound approaches $\approx \frac{1}{s!} - \frac{1}{2(s!)^2}$, which means the probability for the collision is really about $\frac{1}{s!}$.

Proof. We have

$$\frac{1}{n^{s-1}} \binom{q}{s} \leq \frac{q^s}{n^{s-1} s!} = \frac{\alpha^s}{s!} < \frac{1}{s!}$$

Hence upper bound holds. By theorem 2 we have

$$g(n) < \frac{1}{2s!} \frac{q^s}{n^{s-1}} = \frac{\alpha^s}{2s!}$$

$$
\begin{aligned}
f(n) &= e_n^{-(q-s)/n} \\
&> 1 - \frac{q-s}{n}\ln(e_n) \\
&= 1 - \frac{\alpha' \ln(e_n)}{n^{1/s}}
\end{aligned}
$$

The second inequality comes from $e^x > 1 + x, x \neq 0$. Thus,

$$
f(n)(1 - g(n)) \geq f(n) - g(n) > 1 - \frac{\alpha' \ln(e_n)}{n^{1/s}} - \frac{\alpha^s}{2s!}
$$

And

$$
\begin{aligned}
Pr[C(n,q,s)] &\geq f(n)(1-g(n))\frac{1}{n^{s-1}}\binom{q}{s} \\
&> \left(1 - \frac{\alpha' \ln(e_n)}{n^{1/s}} - \frac{\alpha^s}{2s!}\right)\frac{(q-s)^s}{s!} \\
&= \frac{\alpha'^s}{s!} - \left(\frac{\alpha'^{s+1}\ln(e_n)}{s! n^{1/s}} + \frac{\alpha^s \alpha'^s}{2(s!)^2}\right)
\end{aligned}
$$

□

4.2 Tight Bound of $s-$collision

In this section we are going to give a concrete $\Theta(n^{(s-1)/s})$ bound that yields a $s-$collision with probability greater than $1/2$.

Theorem 3
If $2 \leq s \leq q$, and $q = (s!)^{1/s} n^{(s-1)/s} + s - 1 (< n)$, then we have

$$
Pr[C(n,q,s)] > \frac{1}{2} - \left(\frac{s!}{n}\right)^{1/s} \ln e_n
$$

Proof. By AM-GM inequality

$$
\begin{aligned}
\left(\begin{array}{c} q-s \\ s \end{array}\right) &= \frac{(q-s)(q-s-1)\ldots(q-2s+1)}{s!} \\
&< \frac{1}{s!}\left(\frac{(q-s)+(q-s-1)+\ldots+(q-2s+1)}{s}\right)^s \\
&= \frac{[q-(3s-1)/2]^s}{s!} = \frac{\left[(s!)^{1/s} n^{(s-1)/s} - (s+1)/2\right]^s}{s!} \\
&= \left[n^{(s-1)/s} - (s!)^{-1/s}(s+1)/2\right]^s
\end{aligned}
$$

Because $s! \leq (\frac{s+1}{2})^s$, we know $(s!)^{-1/s}(s+1)/2 < 1$. Hence

$$
\left(\begin{array}{c} q-s \\ s \end{array}\right) < \left(n^{(s-1)/s} - 1\right)^s
$$

By $(n-1)^{(s-1)/s} > n^{(s-1)/s} - 1$ we have

$$(n-1)^{s-1} > \left(n^{(s-1)/s} - 1\right)^s$$

Combine two ineuality above we have

$$g(n) = \frac{1}{2(n-1)^{s-1}} \left(\begin{array}{c} q-s \\ s \end{array} \right) < \frac{1}{2}$$

Then we bound $f(x)$ as

$$f(n) > e_n^{-(q-s+1)/n} > 1 - \frac{q-s+1}{n} \ln e_n = 1 - \left(\frac{s!}{n}\right)^{1/s} \ln e_n$$

Combine $f(n)$ and $g(n)$ we get the final result as

$$\Pr[C(n,q,s)] \geq f(n)(1-g(n)) \frac{1}{n^{s-1}} \left(\begin{array}{c} q \\ s \end{array} \right) > \frac{1}{2} - \left(\frac{s!}{n}\right)^{1/s} \ln e_n$$

□

5 Near-Hits

This variant discusses the odds where two birthdays within a distance of k consecutive calendar days are viewed as a collision ($k = 0$ is just the canonical birthday paradox). We call it a k-near-hit in the following discussion. This thread of work was discussed by Abramson and Moser (1970)[1].

5.1 Original results

The idea is simple. First denote the probability of such a collision as $N(n,q,k)$, a collision-free birthday set can be got the following way. First shrink the year to only $n - qk - 1$ days. Set the first member's birthday as the first day of the year, and choose birthdays of the others $\{x_i\}_{2 \leq i \leq n}$ without direct hit. Finally add gaps so that$y_i = x_i + (i-1)k$. The total number of collision-free arragements is

$$\begin{aligned} N(n,q,k) &= \left(\begin{array}{c} n-qk-1 \\ q-1 \end{array} \right) (q-1)! \\ &= \frac{(n-qk-1)!}{(n-(k+1))!} \end{aligned}$$

In this way, the probability of a collision is

$$\begin{aligned} p(n,q,k) &= 1 - \frac{(n-qk-1)!}{(n-q(k+1))!n^{q-1}} \\ &\geq 1 - \prod_{i=1}^{n-1} e^{-\frac{qk+i}{n}} \\ &= 1 - e^{-q(q-1)\frac{k+\frac{1}{2}}{n}} \end{aligned}$$

and a quite accurate approximation to the smallest n for a collision to happen is then (see also Diaconis and Mosteller (1989)[2])

Theorem .1. *When the number of samples are over*

$$q_{col} \lessapprox \frac{1}{2} + \sqrt{2\ln 2 \frac{n}{2k+1} + \frac{1}{4}} \approx \sqrt{2\ln 2}\sqrt{\frac{n}{2k+1}}$$

the probability of a k-near-hit is above $\frac{1}{2}$

And the the previous 'hit' result is a special case of this. see Diaconis and Mosteller (1989)

From the point of view of coincidences, the case $k = 1$ may be the most interesting. In this case, the minimum required number for a 1-near-hit to occur with 0.5 probability is

$$q_{col} \approx 0.68\sqrt{n}$$

which is about $\frac{1}{\sqrt{3}}$ of that for a direct hit.

5.2 Generalization to continuous space

In continuous space there are no direct hits. However, the notion of Near-Hits can be generalized to continuous range, say $[0, 1]$, where two samples within a distance of $\delta \in (0, 1)$ can be regarded as a hit. This case is trivial to analyse, as the probability of no collision after q samples is

$$\Pr[q_{col} > q] = \prod_{i=1}^{q-1} (1 - i\delta)$$

when δ is infinitesimal, the expression above can be bounded by

$$\Pr[q_{col} > q] \leq e^{-\frac{q(q-1)}{2}\delta}$$

$$\inf_q \left\{ \Pr[q_{col} \leq q] \geq \frac{1}{2} \right\} \leq \sqrt{\frac{\ln 2}{\delta}} + 1$$

which is essentially the same as in the discrete case canonical birthday paradox

6 Non-uniform Distribution

Camarri and Pitman (2000)[3]gives an theoretical result for non-uniform distribution. The general case is as follows. (Proof of it will be shown in later parts). The original work discussed the general situation of arbitrary number of collisions. However, in field of cryptography we are more interested in the first collision.

6.1 Main theorem and conclusions

The proof of this theorem is left to Section 6.3

Theorem .2. *, as $n \to \infty$, given a series of distribution on the (asymptotically infinitely large) birthday set $S = \{1, 2, \ldots, n\}$, and probability distributions over it $p : S \to [0, 1]$ with infinitely small converging value $\max_i p_i \to 0$, the probability that no collision happens within q samples is aysmptotically:*

$$\lim_{n\to\infty} \Pr\left[q_{col} > q\left(= \frac{r}{s}\right)\right] = e^{-\frac{r^2}{2}\left(1-\sum \bar{\theta}_i^2\right)} \prod_{i=1}^{n} \left(1 + \bar{\theta}_i r\right) e^{-\bar{\theta}_i r}$$

where

$$\begin{aligned} s &= \sqrt{\sum_{i=1}^{n} p_i^2} \\ \theta_i &= \frac{p_i}{s} \\ \bar{\theta}_i &= \lim_{n\to\infty} \theta_i \\ q &= \frac{r}{s} \end{aligned}$$

Here r is just q being regularized, by a scale of s, to a constant range. θ_i then is also p regularized in a reversed mannor, satisfying $\sum_i \theta_i^2 = 1$. Note here that $\bar{\theta}_i$ is a limit value, and maintains constant in the main equation. Here $\sum_{i\geq 1} \bar{\theta}_i^2 \leq 1$, but the equality does not necessarily hold. More often than not, $\bar{\theta}_i = 0$ for most i.

From this theorem the following corollaries immediately follows:

Corollary .1. *The required number of samples to guarantee a $\frac{1}{2}$ collision probability is*

$$\inf_q \left\{\Pr\left[q_{col} > q\right] \geq \frac{1}{2}\right\} \lesssim \frac{1}{s}\sqrt{\frac{2\ln 2}{\left(1 - \sum_i \bar{\theta}_i^2\right)}}$$

A brief proof of this is as follows:

Proof. First note that when $x > 0$,

$$1 + x \leq e^x$$

The limit probability can then be bounded by

$$\lim_{n\to\infty} \Pr\left[q_{col} > \frac{r}{s}\right] = e^{-\frac{1-\sum \bar{\theta}_i^2}{2} r^2}$$

So

$$\begin{aligned}\inf_q \left\{\Pr\left[q_{col} > q\right] \geq \frac{1}{2}\right\} &\lesssim \inf_q \left\{qs \geq \sqrt{\frac{2\ln 2}{1-\sum_i \bar{\theta}_i^2}}\right\} \\ &= \frac{1}{s}\sqrt{\frac{2\ln 2}{(1-\sum_i \bar{\theta}_i^2)}}\end{aligned}$$

□

In more common situations, when p is not too biased, a simpler result is:

Corollary .2. *Following the notations above, if the probabilities are more evenly distributed by satisfying* $\max_i \theta_i = 0$*,*

$$\lim_{n\to\infty} \Pr\left[q_{col} > \frac{r}{s}\right] = e^{-\frac{r^2}{2}}$$

$$\inf_q \left\{\Pr\left[q_{col} > q\right] \geq \frac{1}{2}\right\} \lesssim \frac{\sqrt{2\ln 2}}{s}$$

And the canonical uniform birthday paradox is just a special case of this.

This is abit wierd at first glance that, when the distribution is quite biased so that $\exists i, \bar{\theta}_i \neq 0$, there is an additional $\frac{1}{\sqrt{1-\sum_i \bar{\theta}_i^2}}$ term which seems to enlarge q_{col}. We have to note here, however, when $\bar{\theta}_i \neq 0$, the corresponding s will be very big and counteract the effect of $\frac{1}{\sqrt{1-\sum_i \bar{\theta}_i^2}}$, and the actual q_{col} should be smaller when p is more biased, which will be shown below in subsection 6.2.

A more general discussion considering the m-th collision and s'-way collisions is given in Subsection 6.4, which gives the following conclusions:

Corollary .3. *Denote* q_{m-col} *as the time step of* m*-th collision, and* $q_{s'-way}$ *as the time step of the first* s'*-way collision, then*

$$\inf_q \left\{\Pr\left[q_m > q\right] \leq \frac{1}{2}\right\} = \mathcal{O}\left(\frac{\sqrt{m}}{s}\right)$$

$$\inf_q \left\{\Pr\left\{q_{s'-way} > q\right\} \leq \frac{1}{2}\right\} = \mathcal{O}\left(\frac{Ns'}{s\sum \bar{\theta}_i}\right)$$

where $N = \left|\left\{i \mid \bar{\theta}_i \neq 0\right\}\right|$ *is the number of non-zero* $\bar{\theta}_i$*'s.*

6.2 Analysis: How to avoid collisions

Conclusion. To minimize the probability of sample collisions, the distribution should be uniform.

Proof. Suppose A is a subset in $N = \{1, \ldots, n\}$, and $N^q = \{A \subset N \mid |A| = q\}$ is set of subsets. Also denote here that $\prod_A = \prod_{i \in A} p_i$. After q steps. Every no-collision sequence of samples is a permutation for some $A \in N^q$. As there are $q!$ permutations for each of A, the probability that no collision happens is

$$\begin{aligned}
\Pr\left[q_{col} > q\right] &= 1 - q! \sum_{A \subset N^q} \prod_A \\
\frac{\partial \Pr}{\partial p_i} &= -q! \sum_{i \in A \subset N^q} \prod_{A-i} \\
&= -q! \sum_{A \subset (N-i)^{q-1}} \prod_A p_i
\end{aligned}$$

Here we relaxed the constraint that $\sum_i p_i = 1$. After projecting $\nabla_p \Pr$ onto the hyperplane, the projected gradient is zero iff the original gradient is orthogonal to the hyperplane, that is, in this case, $\frac{\partial \Pr}{\partial p_i} = \frac{\partial \Pr}{\partial p_j}$ for all $i \neq j$. This then implies $\forall i \neq j$

$$\begin{aligned}
\sum_{A \subset (N-i)^{q-1}} \prod_N p_i &= \sum_{A \subset (N-j)^{q-1}} \prod_A p_j \\
\Rightarrow \sum_{A \subset (N-i-j)^{q-1}} \prod_A (p_i - p_j) &= 0 \\
\Rightarrow (p_i - p_j) &= 0
\end{aligned}$$

The only set of solutions is $p_i = \frac{1}{n}, \forall i$, and would minimize the collision probability. □

6.3 Proof of the theorem

Proof. Consider a Poisson process on $[0, +\infty)$ with rate 1, where events are i.i.d. variables in S following distribution p. Then each random event sequence of any $i \in S$ is an independent Poisson process with rate p_i. There are no collisions iff there are no more than 1 event in each of the sequence.

Within time q, such probability is then

$$\begin{aligned}\Pr\{T_{col} > q\} &= \prod_i \Pr\{N_i \leq 1\} \\ &= \prod_i (1 + p_i q) e^{-p_i q} \\ &= \prod_i (1 + \theta_i r) e^{-\theta_i r} \\ \log \Pr\{T_{col} > q\} &= -\frac{r^2}{2} \sum_i \theta_i^2 + \frac{r^3}{3} \sum_i \theta_i^3 \cdots\end{aligned}$$

if denoting $r = qs$ so that r is scaled to constant value. Now denote the max value $\theta_{max} = \max_i \theta_i$, when $t \in \left(0, \theta_{max}^{-1}\right)$, this log probability can be bounded by □

Lemma .1. *Bounds for the probability above:*

$$\begin{aligned}|\log \Pr\{T_{col} > q\}| &\leq r^2 \frac{\theta_{max}}{2} \left| \sum \theta_i \sum_{j \geq 2} r^{2-j} \theta_{max}^{2-j} \right| \\ &\leq \frac{r^2 \theta_{max}}{2(1 - r\theta_{max})} \sum \theta_i \\ \left| \log \Pr\{T_{col} > q\} + \frac{r^2}{2} \sum_i \theta_i^2 \right| &\leq r^3 \frac{\theta_{max}}{3} \left| \sum \theta_i^2 \sum_{j \geq 2} r^{2-j} \theta_{max}^{2-j} \right| \\ &\leq \frac{r^3 \theta_{max}}{3(1 - r\theta_{max})}\end{aligned}$$

With this, the following lemma showing the limiting property of q_{col} can be derived.

Lemma .2. *If* $\max_i p_i \to 0$*, and let* $T(n)$ *denote the time of the* n*-th Poisson event, then*

$$\frac{q_{col}}{T(q_{col} + 1)} \xrightarrow{P} 1$$

as $n \to \infty$

A brief proof of this is as follows: because of strong law of large numbers, in our Poisson process $\frac{n}{T(n+1)} \xrightarrow{a.s.} 1$ as $T(n+1) \to \infty$. It then suffices to show that $T(q_{col} + 1)$ converges to infinity with $n \to \infty$. This follows from that, as $\theta_{max} \to 0$, $\forall t \in \left(0, \theta_{max}^{-1}\right)$, $|\log \Pr\{T_{col} > q\}| \xrightarrow{P} 0$

Finally, suppose θ_i's are ranked in decreasing manner, for any fixed r set j_r and n_r so that when $n > n_r$, $\theta_{j_r} < \frac{1}{r}$. Dividing the product by j_r, and

$$\lim_{n\to\infty} \prod_{i\le j_r} (1+\theta_i r)\, e^{-\theta_i r} = \prod_{i\le j_r} \left(1+\bar{\theta}_i r\right) e^{-\bar{\theta}_i r}$$

$$\begin{aligned}\log\left(\prod_{i>j_r} (1+\theta_i r)\, e^{-\theta_i r}\right) &= -\sum_{i>j_r} \theta_i^2 \frac{r^2}{2} + \sum_{i>j_r} \theta_i^3 \frac{r^3}{3} \cdots \\ &= -\left(1 - \sum_{i\le j_r} \theta_i^2\right) \frac{r^2}{2} + \sum_{i>j_r} \theta_i^3 \frac{r^3}{3} \cdots\end{aligned}$$

Note that both q and j_r are finite, then$\sum_{i\le j_r} \theta_i^2 \to \sum_{i\le j_r} \bar{\theta}_i^2$ and $\sum_{i\ge j_r} \theta_i^m \to \sum_{i\ge j_r} \bar{\theta}_i^m$ when $m > 2$. The latter one is becaue$\theta_i^m \le \theta_i^2 \theta_{j_r}^{m-2}$, the sum $\sum_{i\ge j_r} \theta_i^m$ achieves uniform convergence as $\forall N > j_r, \forall n$

$$\begin{aligned}\sum_{i\ge N} \theta_i^m &\le \theta_N^{m-2} \sum_{i\ge N} \theta_i^2 \\ &\le \theta_N^{m-2} \\ &\le \max_n \theta_N^{m-2}\end{aligned}$$

which is decreasing as θ_i themselves are decreasing. If, however,

$$\lim_{i\to\infty} \max_n \theta_i = c > 0$$

then there $\exists n$ that $\sum_i \theta_i^2 > 1$. From uniform convergence the exchangability of limiting and infinite summation follows.

Finally summing the two summations gives the result

$$\begin{aligned}\lim_{n\to\infty} \Pr\left\{T_{col} > \frac{r}{s}\right\} &= \lim_{n\to\infty} \prod_i (1+\theta_i r)\, e^{-\theta_i r} \\ &= e^{-\frac{1}{2}\left(1-\sum \bar{\theta}_i^2\right) r^2} \prod_i \left(1+\bar{\theta}_i r\right) e^{-\bar{\theta}_i r}\end{aligned}$$

Finally, using Lemma .2 above, we can substitude T_{col} for q_{col}, and get our desired theorem.

6.4 Generalization: m-th collisions and s-way collision

6.4.1 Distribution for further collisions

This subsection will generalize the previous conclusions concerning only the first collision. The same technique of modeling a Poisson process can also be used to as a basic tool for modelling multiple collisions. The original paper gives the following general theorem:

Theorem .3. *As $n \to \infty$, the sequence of collision steps (denoted by $q_1, q_2, \ldots$, seperately) follows a distribution converging to a superposed independent Poisson processes, i.e.,*

$$(sq_1, sq_2, \ldots, sq_m) \xrightarrow{d} (\eta_1, \eta_2, \cdots, \eta_m)$$

where η_i's are arrival times for superposition of $M^*, M_1, M_2, \cdots$. Here M^* has variable rate of $\left(1 - \sum_i \theta_i^2\right) t$ at time t, and M_i has homogeneous rate of θ_i, with its first point removed.

The proof of this theorem consists of discussion in random coalescent trees, and is not put here right now. The general idea is that the process of random sampling and repetition can be modelled by the growth of a coalescent tree. Utilizing mathematical tool in this area, it can be deduced that the collection of random variable just before the collision is i.i.d and actually independent of the final tree. The deduced independence is another great property, because hereby the full process of collisions can be mathematically the same with a superposition of independent Poisson processes, as shown in the theorem above.

This theorem would immediately give the following more general corollary:

6.4.2 m-th collision

Corollary .4. *Let q_m be the number of samples after which the m-th collision occurs, and $p_q(m) = \lim_{n\to\infty} \Pr[q_m > q]$ the probability generating function is*

$$\begin{aligned} f(x) &= \sum_{m \geq 0} p_q(m) x^m \\ &= e^{\frac{r^2}{2}\left(1 - \sum \bar{\theta}_i^2\right)(x-1) - \sum \bar{\theta}_i r} \prod_i \left(\frac{e^{r\bar{\theta}_i x} - 1}{x} + 1 \right) \end{aligned}$$

where $r = qs$ as before. When distributions are not too biased, i.e. $\max_i \bar{\theta}_i = 0$, we have for $\frac{r^2}{2} > 1$

$$\begin{aligned} f(x) &= e^{\frac{r^2}{2}(x-1)} \\ p_q(m) &= \frac{\Gamma\left(m, \frac{r^2}{2}\right)}{(m-1)!} \\ \inf_q \left\{ \Pr[q_m > q] \leq \frac{1}{2} \right\} &= \mathcal{O}\left(\frac{\sqrt{m}}{s} \right) \end{aligned}$$

Proof. For the first variant Poisson process, the probability generative function f_{M^*} satisfies

$$\log f_{M^*}(x) = \int_0^r t\left(1 - \sum \bar{\theta}_i^2\right)(x-1)\,dt$$
$$f_{M^*}(x) = e^{\frac{t^2}{2}\left(1-\sum \bar{\theta}_i^2\right)(x-1)}$$

and the other processes has function

$$f_{M_i}(x) = e^{-r\bar{\theta}_i}\left(\frac{e^{r\bar{\theta}_i x}-1}{x}+1\right)$$

because their first points are removed.

The expression of $p_q(m)$ in the case of $\max_i \bar{\theta}_i = 0$ comes from.

$$\begin{aligned} p_q(m) &= \sum_{i=0}^{m-1}\frac{1}{i!}f'(0) = e^{-\frac{r^2}{2}}\sum_{i=0}^{m-1}\frac{1}{i!}\left(\frac{r^2}{2}\right)^i \\ &= \frac{(m-1)! - \gamma\left(m, \frac{r^2}{2}\right)}{(m-1)!} = \frac{\Gamma\left(m, \frac{r^2}{2}\right)}{(m-1)!} \end{aligned}$$

Combined with upper bound of incomplete Γ function:

$$\begin{aligned} \Gamma(a,x) &= \int_x^\infty t^{a-1}e^{-t}dt = e^{-x}\int_0^\infty (t+x)^{a-1}e^{-t}dt \\ &\le e^{-x}x^{a-1}\int_0^\infty e^{t\left(\frac{a-1}{x}-1\right)}dt \\ &\le \frac{e^{-x}x^a}{x+1-a} \\ p_q(m) &\le \frac{e^{-x}x^m}{(m-1)!\,(x+1-m)}\,|_{x=\frac{r^2}{2}} \end{aligned}$$

To bound the solution of $p_q(m)$, we take sample values $\frac{\Gamma\left(m,\frac{m}{2}\right)}{(m-1)!} > \frac{1}{2}$ and $\frac{\Gamma(m,2m)}{(m-1)!} < \frac{1}{2}$, and then get a bound for $p_q(m)$ and q_{col}

$$\begin{aligned} p_q(m) &\le \frac{2e^{-x}x^m}{m!} \\ &\le \frac{2(2m)^m}{m!}e^{-x} \\ \inf_q\left\{\Pr[q_m > q] \le \frac{1}{2}\right\} &\le \frac{1}{s}\sqrt{2\ln\left(\frac{4(2m)^m}{m!}\right)} \\ &\le \frac{1}{s}\sqrt{m(2\ln 2+1)+2\ln 2} = \mathcal{O}\left(\frac{\sqrt{m}}{s}\right) \end{aligned}$$

□

The bound of q_{col} given here, however, is not tight when $m \to \infty$.

6.4.3 s'-way collision

We put conclusion here first:

Corollary .5. *Denote $q_{s'-way}$ as the time of the s'-way collision.*

$$\lim_{n\to\infty} \Pr\left\{q_{s-way} > \frac{r}{s}\right\} = \prod \frac{\Gamma\left(s', \bar{\theta}_i r\right)}{(s'-1)!}$$
$$\inf_q \left\{\Pr\left\{q_{s'-way} > q\right\} \leq \frac{1}{2}\right\} = \mathcal{O}\left(\frac{Ns'}{s\sum \bar{\theta}_i}\right)$$

where $N = \left|\left\{i \mid \bar{\theta}_i \neq 0\right\}\right|$

Proof. To avoid notation collision, we here denote s' as the multiplicity in the s'-way collision, and leave $s = \sqrt{\sum p_i^2}$. Following the previous discussion of Poisson process, a s'-way collission happens in category i with probability

$$\begin{aligned}\Pr\left\{T^i_{s-way} > q\right\} &= \prod_i \sum_{j=0}^{s'-1} (\theta_i r)^j e^{-\theta_i r} \\ &= \frac{\Gamma\left(s', \theta_i r\right)}{(s'-1)!}\end{aligned}$$

and following the same discussion as in 6.3 and in proof of .4, the limiting performance of the altogether s-way collision is (when $s > 2$)

$$\begin{aligned}\lim_{n\to\infty} \ln\left[\prod \frac{\Gamma\left(s', \theta_i r\right)}{(s'-1)!}\right] &= \sum_{i|\bar{\theta}_i\neq 0} \ln \frac{\Gamma\left(s', \bar{\theta}_i r\right)}{(s'-1)!} - \sum_{i|\bar{\theta}_i\neq 0} \lim_{\theta_i\to 0} \frac{1}{(s'-1)!} (\theta_i r)^{s'} \\ &= \sum_{i|\bar{\theta}_i\neq 0} \ln \frac{\Gamma\left(s'+1, \bar{\theta}_i r\right)}{(s'-1)!} \\ \lim_{n\to\infty} \Pr\left\{q_{s-way} > \frac{r}{s}\right\} &= \lim_{n\to\infty} \Pr\left\{T_{s'-way} > q\right\} \\ &= \prod \frac{\Gamma\left(s', \bar{\theta}_i r\right)}{(s'-1)!} \\ &\leq \prod \frac{e^{-\bar{\theta}_i r}\left(\bar{\theta}_i r\right)^{s'}}{(s'-1)!\left(\bar{\theta}_i r + 1 - s'\right)} \leq \frac{e^{-\sum \bar{\theta}_i r}\left(\frac{\sum \bar{\theta}_i r}{N}\right)^{Ns'}}{\left(\frac{1}{2}(s'-1)!\right)^N} \\ &\leq \frac{e^{-\sum \bar{\theta}_i r}\left(2s'\right)^{Ns'}}{\left(\frac{1}{2}(s'-1)!\right)^N} \\ \inf_q \left\{\Pr\left\{q_{s-way} > \frac{r}{s}\right\} \leq \frac{1}{2}\right\} &\leq \frac{\left(N(\ln 2 + 1) + s'(N\ln 2 + 1)\right)}{s\sum \bar{\theta}_i} \\ &\leq \frac{4Ns'\ln 2}{s\sum \bar{\theta}_i}\end{aligned}$$

while noticing that $\bar{\theta}_i r > \frac{1}{2}s'$ as it would be harder to get a s'-way collision than a s'-multiple collision, and $\bar{\theta}_i r \leq 2s'$ for the same reason as in .4. $\square$

This is not a tight bound for q_{col}. Another problem is, when $\forall i$, $\bar{\theta}_i = 0$, this probability of no such collision would be 1. So q should be of order higher than $\mathcal{O}\left(\frac{1}{s}\right)$, which agrees with the previous discussion. It is hard yet to give bounds for q_{col}.

7 Future work

Future investigators may be interested in giving a tighter bound for multiple collision and s'-way collision. Also, when $\max \bar{\theta}_i = 0$, q should be of higher order than $\frac{1}{s}$, which may also be discussed.

8 Appendix

8.1 Proof for Trivial lemma

In the deduction there is a equation as below

$$2\sum_{i=2s}^{q} \binom{i-1}{s-1}\binom{i-s}{s} = \binom{q-s}{s}\binom{q}{s}$$

Proof. Do the induction on q. When $q = 2s$, we have $L.H.S = 2\binom{2s-1}{s-1} = \frac{2s}{s}\binom{2s-1}{s-1} = \binom{2s}{s}$ and $R.H.S = \binom{2s}{s} = L.H.S.$, hence the equation holds. Now suppose the equation holds for $q = k$, when $q = k+1$, the increment on left part is $2\binom{k}{s-1}\binom{k+1-s}{s}$, and the right part increment is $\binom{k+1-s}{s}\binom{k+1}{k+1-s} - \binom{k-s}{s}\binom{k}{k-s}$. We aim to prove that these two are equal and thus finish the proof.

$$
\begin{aligned}
&\binom{k+1-s}{s}\binom{k+1}{k+1-s}-\binom{k-s}{s}\binom{k}{k-s}\\
=&\frac{(k+1-s)!}{s!(k+1-2s)!}\frac{(k+1)!}{s!(k+1-s)!}-\frac{(k-s)!}{(k-2s)!s!}\frac{k!}{s!(k-s)!}\\
=&\frac{(k+1)!}{(s!)^2(k+1-2s)!}-\frac{k!}{(s!)^2(k-2s)!}\\
=&\frac{(k+1)!-k!(k-2s+1)}{(s!)^2(k+1-2s)!}\\
=&\frac{k!(k+1-(k-2s+1))}{(s!)^2(k+1-2s)!}\\
=&2\frac{k!s}{(s!)^2(k+1-2s)!}=2\frac{k!}{(s-1)!(k-s+1)!}\frac{(k-s+1)!}{s!(k+1-2s)!}\\
=&2\binom{k}{s-1}\binom{k+1-s}{s}
\end{aligned}
$$

Also there is a combinatorial simple proof for it as below: transform the right handside into

$$\binom{q}{2s}\binom{2s}{s}$$

which means we need to choose 2s to mark and s of which will be marked A (others B). From another perspective, we can first choose the last marked element i (where $i \geq 2s$), and choose $s-1$ same mark from the $i-1$ elements. also choose s another mark in $i-s$ elements. Since element i would either be A or B. Hence the total number of methods would be $2\binom{i-1}{s-1}\binom{i-s}{s}$ for every i. Summation on it and get the left handside. □

References

[1] Abramson, M., Moser, W., 1970. More birthday surprises. Amer. Math. Monthly 7 (7), 856858

[2] Diaconis, P., Mosteller, F., 1989. Methods for studying coincidences. J. Amer. Statist. Assoc. 8 (4), 408, 853861

[3] Camarri, M., Pitman, J., 2000. Limit distributions and random trees derived from the birthday problem with unequal probabilities. Electron. J. Probab. 5, 118.