On Quantifying the Resistance of Concrete Hash Functions to Generic Multi-Collision Attacks

Somindu C. Ramanna and Palash Sarkar

Applied Statistics Unit,
Indian Statistical Institute,
203, B.T. Road, Kolkata,
India 700108.
email: somindu_r@isical.ac.in, palash@isical.ac.in

June 14, 2010

Abstract

Bellare and Kohno (2004) introduced the notion of balance to quantify the resistance of a hash function h to a generic collision attack. Motivated by their work, we consider the problem of quantifying the resistance of h to a generic multi-collision attack. To this end, we introduce the notion of r-balance $\mu_r(h)$ of h and obtain bounds on the success probability of finding an r-collision in terms of $\mu_r(h)$. These bounds show that for a hash function with m image points, if the number of trials q is $\Theta\left(rm^{\left(\frac{r-1}{r}\right)\mu_r(h)}\right)$, then it is possible to find r-collisions with a significant probability of success. The behaviour of random functions and the expected number of trials to obtain an r-collision is studied. These results extend and complete the earlier results obtained by Bellare and Kohno (2004) for collisions (i.e., $r = 2$). Going beyond their work, we provide a new design criteria to provide quantifiable resistance to generic multi-collision attacks. Further, we make a detailed probabilistic investigation of the variation of r-balance over the set of all functions and obtain support for the view that most functions have r-balance close to one.

1 Introduction

An (n, m)-hash function is a map $h : X \to Y$, where $|X| = n$, $|Y| = m$ and $n > m > 0$. A *collision* for h is a pair of *distinct* points $x, x' \in X$ such that $h(x) = h(x')$. Since $n > m$, collisions necessarily exist. For cryptographic applications, h should be designed such that it is infeasible for a resource-bounded adversary to find a collision for h. Such a function is called *collision resistant*. The notion of a collision has been generalized to that of a multi-collision. An r-way collision (or r-*collision*) consists of r *distinct* domain points $x_1, x_2, \cdots, x_r$ such that, $h(x_1) = h(x_2) = \cdots = h(x_r)$. Again, for certain cryptographic applications, the design goal is to ensure that for some suitable range of r, r-collisions are hard to find for a resource-bounded adversary.

Given a hash function h, an algorithm to find an r-collision for h is called an attack. A generic attack does not consider the manner in which the function h is defined, i.e., it does not consider the "internal structure" of h. Instead, some points are picked from the domain and h is applied to them with the hope that a subset of the points will yield an r-collision. In the context of generic attacks, the number of times h is evaluated is taken to be the resource measure of an adversary.

Suppose that q points $x_1, x_2, \cdots, x_q$ are picked. Then the probability of obtaining an r-collision increases monotonically with q. The domain points on which to apply h can be chosen in different ways.

1. **Sampling without replacement.** An r-collision by definition requires the domain points to be distinct. Hence, one would like to use uniform random sampling without replacement to select the domain points. In particular, x_i is selected uniformly at random from $X \setminus \{x_1, \ldots, x_{i-1}\}$. Since it has to be ensured that x_i is distinct from $x_1, \ldots, x_{i-1}$, this method is not very convenient to implement. Also, the lack of independence among the x_i's makes it more difficult to analyse this scenario.

2. **Sampling with replacement.** In this method the domain points are independent and uniformly distributed, i.e., x_i is distributed uniformly over X and is independent of the previous choices. From an algorithmic point of view, this is much more simpler to implement than sampling without replacement.

3. **Picking distinct points without sampling.** Suppose that h is a uniform random function from X to Y. Then it is pointless to use a sampling strategy for picking the domain points. One can simply pick any q distinct points, apply h to them and look for a collision. The probability of success does not depend on the particular set of q points that has been picked. This can also be considered to be the uniform random distribution of q balls to m bins and then looking for a bin with at least r balls.

   In this formulation, the problem has been studied in the literature. McKinney [McK66] gives an exact formula for the probability of finding r-collisions in q trials. But this formula gets more difficult to evaluate as r grows. One can also express this probability using a multinomial cumulative distribution function. Levin [Lev81] provides an efficient way to compute a multinomial distribution function by expressing it as the conditional distribution of independent Poisson random variables given fixed sum. These approximations, however, provide little intuition on the asymptotic behaviour of the complexity of finding an r-collision. For $r = 2$, the complexity is $\Theta(m^{1/2})$ and the attack is usually called the *birthday attack*.

Most works in the cryptography literature follow Point 3 above, i.e., these works ignore the actual hash function and instead analyse a random function. See for example the exposition in [Pre93] and the more recent consideration of the problem in [STKT06]. It is then (implicitly) implied that the results for a random function also hold for the actual hash function.

This approach has been eloquently criticised by Bellare and Kohno [BK04]. They argue that, given a concrete hash function h, one cannot assume that h has "random behaviour", since then, one ends up "not analysing the given h, but rather analysing an abstract and ideal object which ultimately has no connection to h, regardless of the design principle underlying h".

The specific case of $r = 2$ (i.e., collisions) is considered in [BK04]. Suppose that the domain points $x_1, \ldots, x_q$ are chosen using sampling with replacements as explained above. Then, it is usually assumed that the birthday attack applies to the hash function h. Bellare and Kohno [BK04] explain the drawback of this argument. Suppose that a point x is drawn uniformly at random from X. Then it does not follow that the point $h(x)$ is uniformly distributed over Y. Instead, the probability that a $h(x)$ equals a particular $y \in Y$ is $|h^{-1}(y)|/|X|$, where $h^{-1}(y)$ is the set of all pre-images of y under h. So the points $h(x_1), \ldots, h(x_q)$ are uniformly distributed over Y if and only if h is *regular*, i.e., every range point has the same number of pre-images under h. This need not be true for the particular hash function under consideration. In fact, Bellare and Kohno [BK04] comprehensively cover textbook discussions of birthday attacks on hash functions and point out the inadequate and sometimes incorrect viewpoints that have been provided.

Having exposed the fallacy in the analysis of collision resistance of a *concrete* hash function h, Bellare and Kohno [BK04] turn to the problem of quantifying the collision resistance of h. They introduce an important measure $\mu(h)$, called the *balance* of a hash function h. This is defined to be $\mu(h) = -\log_m((n_1^2 + \cdots + n_m^2)/n^2)$, where $Y = \{y_1, \ldots, y_m\}$ and n_i is the number of pre-images of y_i. In other words, $-\mu(h)$ is the logarithm of the probability that $h(x) = h(x')$ for x, x' picked uniformly and independently from X. Note that this includes the possibility that $x = x'$ which is a trivial collision, i.e., $-\mu(h)$ is the logarithm of the probability of obtaining a possibly trivial collision. The rationale for considering possibly trivial collisions in the definition of balance is that if n is large, then with high probability it is a proper collision.

An extensive analysis is carried out to quantify the collision resistance of h in terms of the balance. To this end, two quantities are introduced: $C_h(q)$ and $Q_h(c)$, where $C_h(q)$ is the probability of finding a collision in q trials and $Q_h(c) = \min\{q : C_h(q) \geq c\}$ is the minimum number of queries required to find a collision with probability c. Bounds on $C_h(q)$ are obtained in terms of the balance $\mu(h)$ and these bounds are then translated to obtain bounds on $Q_h(c)$. Section 1.3 summarizes the bounds that they obtain. They further show that regular functions offer (slightly) better collision resistance compared to random functions.

1.1 Our Contributions

The work done by Bellare and Kohno in [BK04] is for $r = 2$. We continue and to a certain extent complete the work started in [BK04] by considering r-collisions for arbitrary $r \geq 2$. As noted above, like [BK04], we also work in the setting where the domain points are chosen according to uniform random sampling with replacement. We call this the generic multi-collision attack. The first question that we consider is the following.

> What is the notion of balance of an (n, m)-hash function h in the context of r-collisions?

To answer this question, we introduce $\mu_r(h)$ which we call the r-balance of the function h. This is defined to be $-(\log_m p_r)/(r-1)$, where p_r is the probability that r points chosen independently and uniformly at random from the domain form an r-collision. For $r_1 < r_2$, we show the relation between r_1-balance and r_2-balance. As in [BK04], the notion of r-balance then leads to the following question.

> How is the performance of the generic multi-collision attack for finding r-collisions related to the notion of r-balance?

Similar to [BK04], we study two quantities.

1. $C_h^{(r)}(q)$. This is the probability of finding an r-collision in q trials.
2. $Q_h^{(r)}(c)$. This is the minimum number of queries required to find an r-collision with probability c.

Upper and lower bounds are obtained on $C_h^{(r)}(q)$. These bounds on $C_h^{(r)}(q)$ are translated to obtain upper and lower bounds on $Q_h^{(r)}(c)$. From this it follows that for an (n, m)-hash function, the number of queries required to find an r-collision with significant probability is $\Theta(rm^{\frac{r-1}{r}\mu_r(h)})$.

Following the agenda set out in [BK04], we next consider a uniform random (n, m)-hash function and introduce $C_{n,m}^{\$(r)}(q)$ (resp. $Q_{n,m}^{\$(r)}(c)$), which is the probability (resp. number of queries) for finding an r-collision with q queries (resp. probability c). Again bounds on $C_{n,m}^{\$(r)}(q)$ are obtained which are used to obtain bounds on $Q_{n,m}^{\$(r)}(c)$. It is shown that if h is a regular (n, m)-hash function, then for a certain range of q, the upper bound on $C_h^{(r)}(q)$ is lesser than a lower bound on $C_{n,m}^{\$(r)}(q)$. As a consequence, using the same number of queries, the probability of finding an r-collision for a regular function is lesser than that of a uniform random function. This shows that compared to random functions, regular functions provide better resistance to the generic multi-collision attack.

Expected number of trials. In Section 4, we provide bounds on the expected number of trials to obtain an r-collision. For collisions, this was done by Bellare and Kohno and we adapt their general arguments to combine with the bounds obtained in this paper.

In an earlier work, Klamkin and Newman [KN67] consider the following problem: given m equally likely alternatives, repeatedly choose the alternatives one by one with replacements until one item occurs r times. They study the expected number of trials for this event to occur and show that as m goes to infinity the expected number of trials is approximately $r\Gamma(1+1/r)m^{(r-1)/r}$, where Γ denotes the usual Gamma function defined by

$$\Gamma(u) = \int_0^\infty e^{-x}x^{u-1}\mathrm{d}x.$$

In Section 4.2, we show the relation between this problem and finding r-collisions for a concrete hash function. In the process, we generalise their approach to work when the alternatives are not necessarily equally likely.

Textbook discussion. Most textbooks analyse collisions obtained by the birthday attack. As mentioned earlier, inadequacies of such analysis has been discussed in [BK04]. On the other hand, to the best of our knowledge, no textbook analyses r-collisions with respect to the generic multi-collision attack. The only analysis available in the literature is using the "balls and bins" approach as discussed above.

Relation to the work of Bellare and Kohno [BK04]. At a general level, we follow the path set out in [BK04]. Some of the results that we obtain for general r have, in a way, been already anticipated by the results for $r = 2$ in [BK04]. Having said this, we would also like to note that our analysis and proofs are not straightforward extensions of [BK04]. Some of the important differences are noted below.

Definition of balance. A straightforward extension of the Bellare and Kohno's definition of balance will be based on the logarithm of $(n_1^r + \cdots + n_m^r)/n^r$. The quantity $(n_1^r + \cdots + n_m^r)/n^r$ is the probability that $h(x_1) = \cdots = h(x_r)$ when $x_1, \ldots, x_r$ are sampled with replacement from the domain. This would include possibly trivial r-collisions, i.e., it would include the possibility that $x_i = x_j$ for some $i \neq j$.

The definition of r-balance that we define is based on the probability of actual r-collisions and not possibly trivial r-collisions. As we show later, this probability is $((n_1)_r + \cdots + (n_m)_r)/n^r$, where $(n_i)_r = n_i(n_i-1)\cdots(n_i-r+1)$. This expression is somewhat more complicated, but, we are able to satisfactorily analyse it. The advantage is that our bounds are better than what would be obtained otherwise.

Lower bound on the success probability. In [BK04], the lower bound on $C_h(q)$ is shown to hold only for a certain range of q.

In contrast, the lower bound on $C_h^{(r)}(q)$ that we obtain holds for all q. This is a consequence of the fact that $C_h^{(r)}(q)$ is monotone increasing in q. (Similarly, $C_h(q)$ is also monotone increasing in q, but, [BK04] do not consider the consequences of this fact.)

Upper bound on the number of queries. The lower bound on success probability translates into an upper bound on the number of queries.

We note an issue of interpretation. In [BK04], it is mentioned that the bounds on $Q_h(c)$ are meaningful only for a certain range of c. But, more precisely, as we point out later, the lower bound on $Q_h^{(r)}(c)$ holds for all c, while the upper bound holds only for a certain range of c. This means that for a value of c outside this range, we cannot upper bound the number of queries required to obtain success probability c. But, we still can say that at least a certain number of queries will be required to obtain success probability c.

Going beyond the Bellare-Kohno agenda. There are two main issues that are considered in this work but have not been considered in [BK04].

1. One criticism about the notion of r-balance is that it is impractical to compute its value for practical hash functions. This may be considered to limit the usefulness of the notion. Our argument against this is twofold.

   First, the notion of r-balance helps in exactly pinning down the resistance of a hash function to generic multi-collision attack. This highlights its central role in our understanding of multi-collisions which is important irrespective of whether one can compute the value or not.

   Second, and from a more practical point of view, we show that this notion leads to a possibly new design criteria for practical hash functions. Suppose that a designer wishes to provably ensure a certain degree of resistance to r-collision attacks, i.e., the designer wishes to prove that finding r-collisions must require a minimum number of hash function evaluations. In Section 2.6, we show that using the notion of r-balance one can satisfactorily give a rather precise answer to this question. In particular, we show that if the number of pre-images of any range point is bounded above, then r-balance has a provable lower bound which translates into a provable lower bound on the number of hash function evaluations to find an r-collision using the generic attack.

2. An important question regarding r-balance is whether most functions have r-balance close to one. We make a detailed investigation of this question using a probabilistic approach. The balance of a random function is a random variable. Probability concentration bounds for this random variable is obtained using Markov inequality, Chebyshev inequality and Chernoff bound. This allows us to support the view that most functions have r-balance close to one.

1.2 Related Work

The property of r-collision freeness has been suggested as a useful tool in building cryptographic protocols. It has been used for the micropayment scheme Micromint of Rivest and Shamir [RS96], for identification schemes by Girault and Stern [GS94] and for signature schemes by Brickell *et. al.* [BPVY00].

The intuition behind relying on r-collision freeness is that finding multi-collisions is harder than finding collisions. This is true when the function is truly random. But concrete hash functions mostly lack "random behaviour". For the case of hash functions based on an iterated construction, Joux [Jou04] has demonstrated that r-collisions in iterated hash functions are not much harder to find than ordinary collisions, even for very large values of r. Following Joux's attack, several works [NS07, HS06] have extended the attack to more general classes of constructions.

There are several space efficient algorithms that find cycles in random graphs. These methods can be used to find collisions in a hash function. It would be interesting to find space efficient algorithms to find multi-collisions. This problem has been addressed recently by Joux and Lucks in [JL09]. They give an algorithm to find 3-collisions that roughly uses m^{δ} storage and whose running time is $m^{1-\delta}$ for $\delta \leq 3$. This shows that finding 3-collisions in time $m^{2/3}$ would require $m^{1/3}$ units of storage.

1.3 Bounds Obtained by Bellare and Kohno [BK04]

The following results summarize the bounds on $C_h(q)$ and $Q_h(c)$ obtained in [BK04].

Theorem 1.1. *[BK04] Let h be an (n, m)-hash function and $m \geq 2$. Let $\alpha \geq 0$ be any real number. Then for any integer $q \geq 2$*

$$(1 - \alpha^2/4 - \alpha) \cdot \binom{q}{2} \cdot \left(\frac{1}{m^{\mu(h)}} - \frac{1}{n}\right) \leq C_h(q) \leq \binom{q}{2} \cdot \left(\frac{1}{m^{\mu(h)}} - \frac{1}{n}\right), \tag{1}$$

the lower bound being true under the additional assumption that

$$q \leq \alpha \cdot \left(1 - \frac{m}{n}\right) \cdot m^{\mu(h)/2}. \tag{2}$$

Theorem 1.2. *[BK04] Let h be an (n, m)-hash function and $n \geq 2m \geq 4$. Let $\alpha \geq 0$ be any real number such that $\beta = 1 - \alpha^2/4 - \alpha > 0$. Let c be a real number in the interval $0 \leq c < 1$. Then*

$$\sqrt{2c} \cdot m^{\mu(h)/2} \leq Q_h(c) \leq 1 + \sqrt{\frac{4c}{\beta}} \cdot m^{\mu(h)/2}, \tag{3}$$

the upper bound being true under the additional assumption that

$$c \leq (\alpha \cdot (1 - m/n) - m^{-\mu(h)/2})^2 \cdot \frac{\beta}{4}. \tag{4}$$

2 Balance-Based Analysis of the Generic Multi-Collision Attack

The *generic multi-collision attack* that we consider is the following. Given an (n, m)-hash function $h : X \to Y$ do the following.

1. Pick $x_1, \ldots, x_q$ independently and uniform at random from X.
2. Compute $y_i = h(x_i)$ for $1 \leq i \leq q$.

An r-collision is found if there are indices $i_1, \ldots, i_r$ with $1 \leq i_1 < i_2 < \cdots < i_r \leq q$ such that $y_{i_1} = \cdots = y_{i_r}$ *and* the domain points $x_{i_1}, \ldots, x_{i_r}$ are distinct. To find an r-collision we certainly need $q \geq r$.

Our goal here is to analyse the performance of the generic multi-collision attack in terms of what we call the r-balance of h. Equivalently, we want to analyse how the following quantities vary with r-balance.

- $C_h^{(r)}(q)$: probability that an r-collision for h is found in q trials ($q \geq r$). This function is monotonically increasing in q since the probability of finding r-collisions cannot decrease as the number of trials increases.
- $Q_h^{(r)}(c)$: the minimum number of trials required to obtain an r-collision with probability greater than or equal to c. That is,
$$Q_h^{(r)}(c) = \min\{q : C_h^{(r)}(q) \geq c\}. \tag{5}$$
Higher the value of c, more is the number of trials needed to find an r-collision. Hence $Q_h^{(r)}(c)$ is monotonically increasing in c.

Note that, for a balance-based analysis of the generic multi-collision attack, the definition of balance given in [BK04] will not be useful. We need to define balance in the context of r-collisions. From the definition, it follows that $C_h^{(2)}(q) = C_h(q)$ and $Q_h^{(2)}(c) = Q_h(c)$.

2.1 Notation

If d is a non-negative integer, then $[d] = \{1, 2, \cdots, d\}$. For an integer $r \geq 2$, $[d]_r$ denotes the set of all r-element subsets of $[d]$. $[d]_{r,2}$ denotes the set of all 2-element subsets of $[d]_r$. Let $r \geq 2$ and $d \geq 0$ be integers. Then $(d)_r$ is defined as follows.

$$(d)_r = \begin{cases} d(d-1)\cdots(d-r+1) & \text{if } d \geq r \\ 0 & \text{otherwise} \end{cases}$$

Let $h : X \to Y$ be an (n, m)-hash function. For any $y \in Y$, $h^{-1}(y) = \{x \in X : h(x) = y\}$. Let $Y = \{y_1, y_2, \cdots, y_m\}$. Then for $i \in [m]$, $n_i = |h^{-1}(y_i)|$ denotes the size of the set of pre-images of y_i under h.

2.2 Definition of r-Balance

A natural way to define the r-balance of h would be in terms of the probability of finding r-collisions for h. To this end, we first prove the following result.

Proposition 2.1. *Let $h : X \to Y$ be a hash function whose domain X and range $Y = \{y_1, y_2, \cdots, y_m\}$ have sizes $n, m \geq r$, respectively. For $i \in [m]$, let $n_i = |h^{-1}(y_i)|$ denote the size of the pre-image of y_i under h. Let r elements be chosen independently and uniformly at random from the domain X. The probability that they form an r-collision is*

$$p_r = \frac{\sum_{i=1}^{m}(n_i)_r}{n^r}.$$

Proof. Let r elements $w_1, w_2, \cdots, w_r$ be picked independently and uniformly at random from the domain X. Let E be the event that these elements form an r-collision. Let A denote the event that these are distinct and for $1 \leq i \leq m$, let B_i be the event that $h(w_1) = \cdots = h(w_r) = y_i$. Then $E = AB_1 \cup AB_2 \cup \cdots \cup AB_m$.

Since B_i's are mutually exclusive events, we have

$$\begin{aligned}
\Pr[E] = \sum_{i=1}^{m} \Pr[AB_i] = \sum_{i=1}^{m} \Pr[A|B_i] \cdot \Pr[B_i] &= \sum_{i=1}^{m} \frac{n_i(n_i-1)\cdots(n_i-r+1)}{n_i^r} \cdot \frac{n_i^r}{n^r} \\
&= \sum_{i=1}^{m} \frac{n_i(n_i-1)\cdots(n_i-r+1)}{n^r}
\end{aligned}$$

Since $p_r = \Pr[E]$, the proposition follows. $\square$

Definition 2.1. Let $h : X \to Y$ be a hash function with $|X| = n$ and $Y = \{y_1, y_2, \cdots, y_m\}$. Let $n \geq r$ and $p_r > 0$. The *r-balance of h*, denoted $\mu_r(h)$, is defined as

$$\mu_r(h) = \frac{1}{r-1} \cdot \log_m \left(\frac{1}{p_r}\right). \tag{6}$$

If $n_i < r$ for all i, then there cannot be any r-collisions, that is, $p_r = 0$. A necessary condition for the existence of an r-collision is that $n_i \geq r$ for at least one i. If $n \geq rm$, then an r-collision will certainly exist but there could be an r-collision even if $n < rm$. We only require the condition that $p_r > 0$.

Consider the case $r = 2$. From the definition of $\mu(h)$, we have

$$m^{-\mu_2(h)} = \frac{\sum_{i=1}^m n_i(n_i-1)}{n^2} = \frac{\sum_{i=1}^m {n_i}^2}{n^2} - \frac{\sum_{i=1}^m n_i}{n} = m^{-\mu(h)} - \frac{1}{n}.$$

This shows that $\mu_2(h)$ is always greater than $\mu(h)$. The difference gets smaller as n grows larger.

The following lemma will be useful in obtaining bounds on the r-balance of a hash function.

Lemma 2.2. *Let $r \geq 2$ be an integer. Let $n_1, n_2, \cdots, n_m$ be non-negative integers such that $\sum_{i=1}^m n_i = n$. Then*

$$m \cdot \left(\frac{n}{m}\right)_r \leq \sum_{i=1}^m (n_i)_r \leq (n)_r.$$

The upper bound is attained when exactly one of the n_i equals n and all others are zero, while the lower bound is attained when all the n_is are equal.

Proof. We will prove the bounds using a counting argument. Let $S(n_i)$ denote the set of all distinct arrangements of n_i things taken r at a time. Then $|S(n_i)| = (n_i)_r$ for $i = 1, \cdots m$. If $n_j \leq r-1$ for some j then $S(n_j) = \emptyset$. Assume, without loss of generality, that the first k of the n_i's are greater than $r-1$. By definition $n = \sum_{i=1}^m n_i$. Let S denote the set of all distinct arrangements of n things taken r at a time. Each arrangement in $S(n_i)$ is also present in S. This shows that $S(n_1) \cup S(n_2) \cup \cdots \cup S(n_k) \subseteq S$. Also since the $S(n_i)$'s are disjoint, we have

$$(n_1)_r + (n_2)_r + \cdots + (n_k)_r \leq (n_1 + n_2 + \cdots + n_k)_r = (n)_r$$

Equality occurs when $k = 0$ i.e., one of the n_i's is equal to n and the rest are zero. This gives the upper bound on $\sum_{i=1}^m (n_i)_r$.

Now we claim that $\sum_{i=1}^m (n_i)_r$ attains its minimum when all n_i's are equal i.e., $n_1 = n_2 = \cdots = n_m = \frac{n}{m}$. Suppose there exist n_i and n_j such that $n_i > \frac{n}{m}$ and $n_j < \frac{n}{m}$. Assume, without loss of generality, that $i = 1$ and $j = 2$. To prove the claim, we need but show that

$$(n_1 - 1)_r + (n_2 + 1)_r + \cdots + (n_k)_r < (n_1)_r + (n_2)_r + \cdots + (n_k)_r.$$

Let T_i denote the set containing n_i items. Clearly, $T_1 \cup T_2 \cup \cdots \cup T_m = X$. Let $x \in T_1$. The number of arrangements of items in T_1 taken r at a time that contain x is equal to $r(n_1 - 1)_{r-1}$. Suppose we remove x from T_1 and put it in T_2. Then the number of arrangements of items in T_2 taken r at a time that contain x is equal to $r(n_2)_{r-1}$. Thus we have

$$\begin{aligned}
&((n_1)_r + (n_2)_r + \cdots + (n_k)_r) - ((n_1-1)_r + (n_2+1)_r + \cdots + (n_k)_r) \\
&= |S(n_1) \cup S(n_2) \cup \cdots \cup S(n_m)| - |S(n_1 - 1) \cup S(n_2 + 1) \cup \cdots \cup S(n_m)| \\
&= |S(n_1) \cup S(n_2)| - |S(n_1 - 1) \cup S(n_2 + 1)| \\
&= |S(n_1 - 1)| + r(n_1 - 1)_{r-1} + |S(n_2)| - |S(n_1 - 1)| - |S(n_2)| - r(n_2)_{r-1} \\
&= r(n_1 - 1)_{r-1} - r(n_2)_{r-1} \\
&> 0
\end{aligned}$$

since $n_1 - 1 > n_2$. This gives the lower bound. □

The following proposition provides the minimum and maximum values of the r-balance of a function and the conditions under which they are attained. The proof follows directly from the definition of $\mu_r(h)$ and Proposition 2.2.

Proposition 2.3. *Let h be an (n,m)-hash function. Then*

$$\frac{1}{r-1}\log_m \frac{n^r}{(n)_r} \le \mu_r(h) \le \frac{1}{r-1}\log_m \frac{n^r}{m \cdot \left(\frac{n}{m}\right)_r} \tag{7}$$

The lower bound is attained when h is a constant function and the upper bound is attained when h is a regular function.

Let $\mu_r^{\min}(n,m)$ and $\mu_r^{\max}(n,m)$ denote the minimum and maximum values of the r-balance of an (n,m)-hash function. The quantity $\mu_r^{\min}(n,m)$ can be approximated as follows.

$$\begin{aligned}\mu_r^{\min}(n,m) = \frac{1}{r-1}\log_m \frac{n^r}{(n)_r} &= \frac{1}{r-1}\log_m \frac{1}{\left(1-\frac{1}{n}\right)\cdots\left(1-\frac{r-1}{n}\right)} \\ &\approx \frac{1}{r-1}\log_m \frac{1}{e^{-1/n}\cdots e^{-(r-1)/n}} = \frac{1}{r-1}\log_m \frac{1}{e^{-\binom{r}{2}/n}} = \frac{r}{2n(\ln m)}.\end{aligned}$$

This shows that, for large n, the $\mu_r^{\min}(n,m)$ is close to zero. Similarly one can approximate $\mu_r^{\max}(n,m)$ as follows.

$$\begin{aligned}\mu_r^{\max}(n,m) = \frac{1}{r-1}\log_m \frac{n^r}{m \cdot \left(\frac{n}{m}\right)_r} &= \frac{1}{r-1}\log_m \frac{m^{r-1}}{\left(1-\frac{m}{n}\right)\cdots\left(1-\frac{(r-1)m}{n}\right)} \\ &\approx \frac{1}{r-1}\log_m \frac{m^{r-1}}{e^{-m/n}\cdots e^{-(r-1)m/n}} \\ &= \frac{1}{r-1}\log_m \left(m^{r-1}e^{\binom{r}{2}m/n}\right) = 1 + \frac{rm}{2n(\ln m)}.\end{aligned}$$

This shows that for large n, $\mu_r^{\max}(n,m)$ is close to one.

2.3 Relation Between r_1-Balance and r_2-Balance

One natural question that arises is whether it is easy to find r_2-collisions for a given hash function given that r_1-collisions can be found easily, for $r_1 \neq r_2$. We will analyse this by looking at the difference between the r_1-balance and r_2-balance of the function. In the following discussion, we will write μ_r in place of $\mu_r(h)$.

Proposition 2.4. *Let h be an (n,m)-hash function with $n_1 \ge n_2 \ge \cdots \ge n_m$ where the n_i's are as defined earlier. Let $2 \le r_1 < r_2 \le n_1$. Let n_j be the smallest among the n_i's which is greater than or equal to r_2. Let the function f_{r_1,r_2} be defined as follows:*

$$f_{r_1,r_2}(x) = (x-r_1)(x-r_1-1)\cdots(x-r_2+1).$$

Then

$$\left(\frac{r_2-1}{r_1-1}\right)\mu_{r_2} - \frac{1}{r_1-1}\log_m \frac{n^{r_2-r_1}}{f_{r_1,r_2}(n_j)} \le \mu_{r_1} \le \left(\frac{r_2-1}{r_1-1}\right)\mu_{r_2} - \frac{1}{r_1-1}\log_m \frac{n^{r_2-r_1}}{f_{r_1,r_2}(n_1)}.$$

Note. For practical values of n, m and small r_2, n_j will be equal to n_m for most functions.

Proof. Let r_1 be fixed. We will prove this result using induction on r_2. For the base case, suppose that $r_2 = r_1 + 1 = r$, say.

$$(n_j - r + 1)\sum_{i=1}^{m}(n_i)_{r-1} \leq \sum_{i=1}^{m}(n_i)_r \leq (n_1 - r + 1)\sum_{i=1}^{m}(n_i)_{r-1}. \tag{8}$$

From the definition of r-balance we have

$$(r-1)\mu_r - (r-2)\mu_{r-1} = \log_m n + \log_m \frac{\sum_{i=1}^{m}(n_i)_{r-1}}{\sum_{i=1}^{m}(n_i)_r}. \tag{9}$$

Combining inequality (8) and Equation (9) we get

$$\log_m \frac{n}{n_1 - r + 1} \leq (r-1)\mu_r - (r-2)\mu_{r-1} \leq \log_m \frac{n}{n_j - r + 1}$$

$$\left(\frac{r-1}{r-2}\right)\mu_r - \frac{1}{r-2}\log_m \frac{n}{n_j - r + 1} \leq \mu_{r-1} \leq \left(\frac{r-1}{r-2}\right)\mu_r - \frac{1}{r-2}\log_m \frac{n}{n_1 - r + 1}. \tag{10}$$

This completes the proof of the base case. Suppose now that the result holds for $r_2 - 1$. We need to show that it holds for r_2. First, consider the lower bound. By induction hypothesis we have

$$\mu_{r_1} \geq \left(\frac{r_2 - 2}{r_1 - 1}\right)\mu_{r_2 - 1} - \frac{1}{r_1 - 1}\log_m \frac{n^{r_2 - r_1 - 1}}{f_{r_1, r_2 - 1}(n_j)}. \tag{11}$$

Inequality (10) gives us a lower bound on μ_{r_2} as follows:

$$\mu_{r_2 - 1} \geq \left(\frac{r_2 - 1}{r_2 - 2}\right)\mu_{r_2} - \frac{1}{r_2 - 2}\log_m \frac{n}{n_j - r_2 + 1}.$$

Substituting this lower bound in inequality (11), we get,

$$\begin{aligned}
\mu_{r_1} &\geq \left(\frac{r_2 - 2}{r_1 - 1}\right)\left(\left(\frac{r_2 - 1}{r_2 - 2}\right)\mu_{r_2} - \frac{1}{r_2 - 2}\log_m \frac{n}{n_j - r_2 + 1}\right) - \frac{1}{r_1 - 1}\log_m \frac{n^{r_2 - r_1 - 1}}{f_{r_1, r_2 - 1}(n_j)} \\
&= \left(\frac{r_2 - 1}{r_1 - 1}\right)\mu_{r_2} - \frac{1}{r_1 - 1}\log_m \frac{n^{r_2 - r_1}}{(n_j - r_2 + 1) f_{r_1, r_2 - 1}(n_j)} \\
&= \left(\frac{r_2 - 1}{r_1 - 1}\right)\mu_{r_2} - \frac{1}{r_1 - 1}\log_m \frac{n^{r_2 - r_1}}{f_{r_1, r_2}(n_j)}.
\end{aligned}$$

Applying the same technique to the upper bound, we get the required result. ∎

The following corollary follows directly from Proposition 2.4 and is simpler to understand.

Corollary 2.5.

$$\left|\mu_{r_1} - \left(\frac{r_2 - 1}{r_1 - 1}\right)\mu_{r_2}\right| < \frac{1}{r_1 - 1}\log_m \frac{n^{r_2 - r_1}}{f_{r_1, r_2}(n_j)} \leq \frac{1}{r_1 - 1}\log_m \frac{n^{r_2 - r_1}}{f_{r_1, r_2}(n_m)}.$$

We can interpret Proposition 2.4 as follows: If r_2 collisions can be found easily for a given hash function, then it is not much harder to find r_1 collisions for $r_2 > r_1$. If r_1-balance (resp. r_2-balance) is known for a function, then one can get some idea of the value r_2-balance (resp. r_1-balance). Note that the bounds coincide for regular and constant functions.

Note. If r_2-collisions do not exist, then we cannot say anything about how hard it is to find r_1-collisions. For example, consider a function for which $(r+1)$-collisions do not exist. Then $n_i \leq r$ for all $i \in \{1, \cdots, m\}$. Finding r-collisions for functions of such form is not necessarily easy. If $n_i = r$ for all i (and so $n = rm$), then the function is a regular function and has the maximum balance. We later show that such functions offer the maximum resistance (among all (rm, m) functions) to generic multi-collision attacks.

2.4 Bounds on $C_h^{(r)}(q)$

For $I \in [q]_r$, $I = \{i_1, i_2, \cdots, i_r\}$, define a random variable Z_I as follows.

$$Z_I = \begin{cases} 1 & \text{if } x_{i_1}, x_{i_2}, \cdots, x_{i_r} \text{ form an } r\text{-collision} \\ 0 & \text{otherwise} \end{cases}$$

From Proposition 2.1 and the definition of r-balance we have

$$\mathrm{E}[Z_I] = \Pr[Z_I = 1] = \frac{\sum_{i=1}^{m}(n_i)_r}{n^r} = m^{-(r-1)\mu_r(h)} = p_r \tag{12}$$

Then $Z = \sum_{I \in [q]_r} Z_I$ denotes the number of r-collisions. The expected value of Z is $\binom{q}{r} m^{-(r-1)\mu_r(h)}$. We are interested in an r-collision and would like to know the number of queries required to have the expected value of Z to be equal to 1. This is given by the value of q such that $(q)_r = r! \times m^{(r-1)\mu_r(h)}$. Using the inequality $(q-r)^r < (q)_r$, it can be easily shown that choosing $q = r + (r!)^{1/r} \times m^{(r-1)\mu_r(h)/r}$ ensures that $\mathrm{E}[Z] \geq 1$. This gives an indication of the "right" value of q required to obtain an r-collision.

We now consider that q trials are made and obtain bounds on $C_h^{(r)}(q)$. An upper bound on $C_h^{(r)}(q)$ is easy to obtain.

Theorem 2.6 (Upper Bound on $C_h^{(r)}(q)$)**.** *Let h be an (n, m)-hash function with $n \geq r$ and $m \geq 2$. Then for any integer $q \geq r$,*

$$C_h^{(r)}(q) \leq \binom{q}{r} p_r. \tag{13}$$

Proof. Let $\{i_1, \ldots, i_r\} \subseteq [q]$. The probability that $x_{i_1}, \ldots, x_{i_r}$ forms an r-collision is p_r. The result now follows from the union bound on probability. □

To obtain a lower bound on $C_h^{(r)}(q)$, we need the following lemma.

Lemma 2.7. *Let h be an (n, m)-hash function and ℓ be an integer such that $\ell > r$. Then*

$$\left(\sum_{i=1}^{m}(n_i)_\ell\right)^r \leq \left(\sum_{i=1}^{m}(n_i)_r\right)^\ell. \tag{14}$$

As a consequence, $p_\ell \leq p_r^{\ell/r}$.

Proof. Without loss of generality assume that $n_1 \geq n_2 \geq \cdots \geq n_m$. Let $A_i = (n_i)_\ell$, $B_i = (n_i)_r$ and $C_i = (n_i - r)\cdots(n_i - \ell + 1)$, so that $A_i = B_i C_i$. We are required to show

$$(B_1C_1 + \cdots + B_mC_m)^r \leq (B_1 + \cdots + B_m)^\ell. \tag{15}$$

Consider the multinomial expansion of the left hand side of this equation. A term of this expansion is of the form

$$\frac{r!}{d_1!\ d_2! \cdots d_m!}(B_1C_1)^{d_1}(B_2C_2)^{d_2}\cdots(B_mC_m)^{d_m}$$

where $d_1 + \cdots + d_m = r$. We show that this term is less than or equal to

$$\frac{\ell!}{(d_1 + \ell - r)!\ d_2! \cdots d_m!} B_1^{d_1+(\ell-r)} B_2^{d_2} \cdots B_m^{d_m}$$

which (since $\ell > r$) is a term in the multinomial expansion of the right hand side of (15). This inequality is shown by separately proving the following two inequalities.

1. $\frac{r!}{d_1!\ d_2!\cdots d_m!} \leq \frac{\ell!}{(d_1+\ell-r)!\ d_2!\cdots d_m!}$.

2. $(B_1C_1)^{d_1}(B_2C_2)^{d_2}\cdots(B_mC_m)^{d_m} \leq B_1^{d_1+(\ell-r)}B_2^{d_2}\cdots B_m^{d_m}$.

Point (1) holds if $\dfrac{r!}{d_1!} \leq \dfrac{\ell!}{(d_1+\ell-r)!}$, i.e., if

$$\frac{\ell(\ell-1)\cdots(r+1)}{(d_1+\ell-r)(d_1+\ell-r-1)\cdots(d_1+1)} \geq 1.$$

This inequality holds if for $1 \leq j \leq \ell-r$, $r+j \geq d_1+j$ which clearly holds since $d_1 \leq r$.

Now consider the second point, which holds if $C_1^{d_1}C_2^{d_2}\cdots C_m^{d_m} \leq B_1^{\ell-r}$. For $1 \leq j \leq \ell-r$, let $E_j = (n_1-r-j+1)^{d_1}\cdots(n_m-r-j+1)^{d_m}$. Clearly $E_j \leq E_1$ for $1 \leq j \leq \ell-r$. Then, it follows that

$$C_1^{d_1}C_2^{d_2}\cdots C_m^{d_m} = E_1E_2\cdots E_{\ell-r} \leq {E_1}^{\ell-r}$$

Point (2) now follows if $E_1 \leq B_1$. Using the assumption that $n_1 \geq n_i$ for $1 \leq i \leq m$, it follows that

$$E_1 = (n_1-r)^{d_1}\cdots(n_m-r)^{d_m} \leq (n_1-r)^{d_1+\cdots+d_m} = (n_1-r)^r \leq (n_1)_r = B_1$$

This completes the proof of (14).

By definition,

$$p_\ell = \frac{\sum_{i=1}^m (n_i)_\ell}{n^\ell} \leq \frac{\left(\sum_{i=1}^m (n_i)_r\right)^{\ell/r}}{n^\ell} = \left(\frac{\sum_{i=1}^m (n_i)_r}{n^r}\right)^{\ell/r} = p_r^{\ell/r}.$$

This completes the proof. □

Theorem 2.8 (Lower Bound on $C_h^{(r)}(q)$)**.** *Let h be an (n,m)-hash function with $n \geq r$ and $m \geq 2$. Then*

$$C_h^{(r)}(q) \geq \frac{1}{2}\left(2 - \sum_{k=0}^{r-1}\binom{r}{k}\binom{q-r}{r-k}{p_r}^{(r-k)/r}\right)\cdot\binom{q}{r}\cdot p_r. \tag{16}$$

Proof. Let $[q]_{r,2}$ denote the set of all 2-element subsets of $[q]_r$. By the principle of inclusion and exclusion, we have

$$\begin{aligned} C_h^{(r)}(q) &= \Pr\left[\bigvee_{I\in[q]_r} Z_I = 1\right] && (17)\\ &= \sum_{I\in[q]_r}\Pr[Z_I=1] - \sum_{\substack{I,J\in[q]_r\\ I\neq J}}\Pr[Z_I=1\wedge Z_J=1]\\ &\quad + \cdots + (-1)^{\binom{q}{r}-1}\Pr\left[\bigwedge_{I\in[q]_r} Z_I=1\right] && (18)\end{aligned}$$

Considering the first two terms in the above equation will give us a lower bound on $C_h^{(r)}(q)$.

$$C_h^{(r)}(q) \geq \sum_{I\in[q]_r}\Pr[Z_I=1] - \sum_{\{I,J\}\in[q]_{r,2}}\Pr[Z_I=1\wedge Z_J=1] \tag{19}$$

$$\sum_{I\in[q]_r} \Pr[Z_I = 1] = \binom{q}{r} \Pr[Z_I = 1] = \binom{q}{r} \cdot p_r \tag{20}$$

In order to obtain the required lower bound, we need to maximize the second term of Equation (19). This is where our proof deviates from the one given in [BK04].

For $k = 0, 1, \cdots, r-1$, let T_k be the number of pairs $\{I, J\} \in [q]_{r,2}$ such that $|I \cap J| = k$. The k common elements can be chosen in $\binom{q}{k}$ ways. The remaining $r-k$ elements in I can be chosen in $\binom{q-k}{r-k}$ ways and for each such I, we can choose the remaining $r-k$ elements in J in $\binom{q-r}{r-k}$ ways. But this way we are counting every unordered pair twice (i.e., $\{I, J\}$ and $\{J, I\}$ are indistinguishable but both are counted). Therefore, we have

$$T_k = \frac{1}{2}\binom{q}{k}\binom{q-k}{r-k}\binom{q-r}{r-k} = \frac{1}{2}\binom{q}{r}\binom{r}{k}\binom{q-r}{r-k} \tag{21}$$

We can now break up the second term in Equation (19) as follows:

$$\sum_{\{I,J\}\in[q]_{r,2}} \Pr[Z_I = 1 \wedge Z_J = 1] = \sum_{k=0}^{r-1} T_k \cdot \Pr[Z_I = 1 \wedge Z_J = 1 \mid (|I \cap J| = k)] \tag{22}$$

Let x_I and x_J denote the set of points corresponding to the index sets I and J respectively. Suppose I and J are disjoint (i.e., $k = 0$). Due to the disjointedness of I and J and the sampling strategy, the value Z_I takes is independent of the value Z_J takes. So when $k = 0$, we have,

$$\Pr[Z_I = 1 \wedge Z_J = 1] = \Pr[Z_I = 1] \cdot \Pr[Z_J = 1] = {p_r}^2. \tag{23}$$

When $k \geq 1$, the events $Z_I = 1$ and $Z_J = 1$ indicate that the elements in x_I map to a common point and so do the elements in x_J. Since $I \cap J \neq \emptyset$, the common image of the elements of both x_I and x_J must be the same. Hence $\Pr[Z_I = 1 \wedge Z_J = 1|(|I \cap J| = k)]$ is the probability that the $2r-k$ distinct elements corresponding to the index set $I \cup J$ form a $2r-k$-collision. That is,

$$\Pr[Z_I = 1 \wedge Z_J = 1] = p_{2r-k} \tag{24}$$

Combining Equations (21), (22), (23) and (24), we obtain the following:

$$\Pr[Z_I = 1 \wedge Z_J = 1] = T_0 \cdot {p_r}^2 + \sum_{k=1}^{r-1} T_k \cdot p_{2r-k} \tag{25}$$

To obtain an upper bound on the above expression, we need an upper bound on p_{2r-k}. From Lemma 2.7, we have

$$p_{2r-k} \leq p_r^{(2r-k)/r} = p_r p_r^{(r-k)/r}. \tag{26}$$

Combining Equations (21), (25) and (26), we obtain

$$\sum_{\{I,J\}\in[q]_{r,2}} \Pr[Z_I = 1 \wedge Z_J = 1] \leq \frac{1}{2}\binom{q}{r} \cdot p_r \cdot \sum_{k=0}^{r-1} \binom{r}{k}\binom{q-r}{r-k} {p_r}^{(r-k)/r} \tag{27}$$

Combining Equations (19), (20) and (27), we obtain the lower bound stated in Equation (16) as follows.

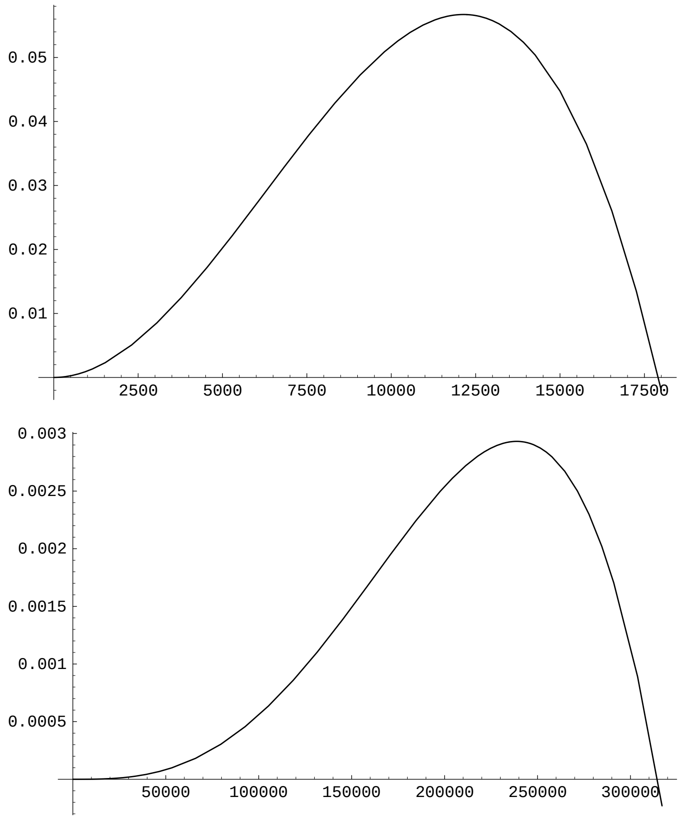

Figure 1: Behaviour of $L_h^{(r)}(q)$ for $r = 2$ and $r = 3$ with $m = 2^{32}$ and $\mu_r = 0.9$.

$$\begin{aligned} C_h^{(r)}(q) &\geq \binom{q}{r} \cdot p_r - \frac{1}{2}\binom{q}{r} \cdot p_r \cdot \sum_{k=0}^{r-1} \binom{r}{k}\binom{q-r}{r-k} {p_r}^{(r-k)/r} \\ &= \frac{1}{2}\binom{q}{r} \cdot p_r \cdot \left(2 - \sum_{k=0}^{r-1} \binom{r}{k}\binom{q-r}{r-k} {p_r}^{(r-k)/r}\right) \end{aligned}$$

□

Towards a better lower bound. We now discuss the behaviour of this lower bound. Let

$$s_h^{(r)}(q) = 2 - \sum_{k=0}^{r-1} \binom{r}{k}\binom{q-r}{r-k} {p_r}^{(r-k)/r}$$

and let the lower bound of Theorem 2.8 be denoted $L_h^{(r)}(q)$. We have

$$L_h^{(r)}(q) = \frac{1}{2} \cdot p_r \binom{q}{r} s_h^{(r)}(q).$$

$L_h^{(r)}(q)$ is a polynomial in q of degree $2r$. One can make the following observations about this polynomial.

Table 1: h is a hash function with $n = 2^{512}$, $m = 2^{160}$ and $\mu_2(h) = 0.8$.

r	$\mathsf{qmax}_h^{(r)}$	Lower bound on $C_h^{(r)}(q)$	Upper bound on $C_h^{(r)}(q)$	Ratio (upper bound/lower bound)
3	1.9327×10^{25}	2.9312×10^{-3}	1.0391×10^{-2}	3.5449
4	2.4385×10^{28}	8.6908×10^{-5}	3.7393×10^{-4}	4.3025
5	1.6855×10^{30}	1.6724×10^{-6}	8.4565×10^{-6}	5.0565
6	2.7482×10^{31}	2.2583×10^{-8}	1.3117×10^{-7}	5.8083
7	1.9718×10^{32}	2.2585×10^{-10}	1.4813×10^{-9}	6.5587
8	8.4939×10^{32}	1.7403×10^{-12}	1.2719×10^{-11}	7.3085
9	2.6088×10^{33}	1.065×10^{-14}	8.5817×10^{-14}	8.0579
10	6.3321×10^{33}	5.3018×10^{-17}	4.6689×10^{-16}	8.8062

- The binomial coefficient $\binom{q}{r} = \frac{1}{r!}q(q-1)\cdots(q-(r-1))$ and it vanishes at the points $0, 1, \cdots, r-1$ which means these are roots of $L_h^{(r)}(q)$. It is also monotone increasing and positive for $q \geq r$.
- $s_h^{(r)}(q)$ is decreasing in q and becomes negative after a certain point causing $L_h^{(r)}(q)$ to decrease.
- The polynomial $s_h^{(r)}(q)$ has exactly one sign change and by Descartes' rule of signs it will have at most one positive real root.

These observations show that $L_h^{(r)}(q)$ has exactly $r+1$ non-negative real roots including $0, 1, \cdots, r-1$. This is because $L_h^{(r)}(q)$ is positive at $q = r$ and after a certain point becomes negative which means it is zero at exactly one point after $r-1$. Let the $(r+1)^{st}$ real root be denoted as θ. In the interval ranging from $q = r$ to $q = \theta$, the curve representing $L_h^{(r)}(q)$ must have one turning point. Figure 1 gives some examples to show how $L_h^{(r)}(q)$ behaves. Let the value of q at which the curve turns be denoted $\mathsf{qmax}_h^{(r)}$ and let $\mathsf{cmax}_h^{(r)} = L_h^{(r)}(\mathsf{qmax}_h^{(r)})$. For $q \leq \mathsf{qmax}_h^{(r)}$ the lower bound will be $L_h^{(r)}(q)$. For $q > \mathsf{qmax}_h^{(r)}$, $L_h^{(r)}(q)$ is decreasing but the probability of finding r-collisions cannot decrease as we increase the number of trials. Hence $L_h^{(r)}(\mathsf{qmax}_h^{(r)})$ is a better lower bound. Based on this discussion and Theorems 2.6 and 2.8, we are able to state more appropriate bounds on $C_h^{(r)}(q)$.

Theorem 2.9. *Let h be an (n, m)-hash function. Then*

$$\max_{r \leq t \leq q} L_h^{(r)}(t) \leq C_h^{(r)}(q) \leq \binom{q}{r} \cdot p_r. \tag{28}$$

Note. Theorem 2.9 is valid for all q (and for all $r \geq 2$). This is to be contrasted with the bound obtained in [BK04] for the case $r = 2$ (see Theorem 1.1).

How close are the bounds? Since $L_h^{(r)}(q)$ is difficult to analyse, we provide computational results to show how close the bounds are. Table 1 provides lower and upper bounds on $C_h^{(r)}(q)$ for different values of r and a fixed h. Both the bounds are evaluated at $\mathsf{qmax}_h^{(r)}$. For values of $q \geq \mathsf{qmax}_h^{(r)}$, the gap between the bounds increases. The table indicates that the bounds are quite close. The ratio increases by around 0.75 with every one-step increases in r.

The lower bound stated in Theorem 2.9 can be further simplified as shown below.

Corollary 2.10. *Let h be an (n,m)-hash function. Assume $n \geq r \geq 2$. Let*

$$\alpha(q) = qm^{-(\frac{r-1}{r})\mu_r(h)}. \tag{29}$$

Then $C_h^{(r)}(q) \geq \max_{r \leq t \leq q} \frac{1}{2} \left(3 - (\alpha(t)+1)^r\right) \cdot \binom{t}{r} \cdot m^{-(r-1)\mu_r(h)}$.

Proof. We proceed as in the proof of Theorem 2.8 upto Equation (27). It is after this point that the proof will deviate. Using Equations (19), (20) and (27) we get

$$\begin{aligned}
C_h^{(r)}(q) &\geq \binom{q}{r} \cdot p_r - \frac{1}{2}\binom{q}{r} \cdot p_r \cdot \sum_{k=0}^{r-1} \binom{r}{k}\binom{q-r}{r-k} {p_r}^{(r-k)/r} \\
&= \frac{1}{2}\binom{q}{r} \cdot p_r \cdot \left(2 - \sum_{k=0}^{r-1} \binom{r}{k}\binom{q-r}{r-k} {p_r}^{(r-k)/r}\right) \\
&\geq \frac{1}{2}\binom{q}{r} \cdot p_r \cdot \left(2 - \sum_{k=0}^{r-1} \binom{r}{k} q^{r-k} {p_r}^{(r-k)/r}\right) \\
&= \frac{1}{2}\binom{q}{r} \cdot p_r \cdot \left(2 - \sum_{k=0}^{r-1} \binom{r}{k} (\alpha(q))^{r-k}\right) \\
&= \frac{1}{2}\binom{q}{r} \cdot p_r \cdot \left(2 - ((\alpha(q)+1)^r - 1)\right) \\
&= \frac{1}{2}\binom{q}{r} \cdot p_r \cdot \left(3 - (\alpha(q)+1)^r\right).
\end{aligned}$$

Using the same arguments that led to Theorem 2.9, we get the bound stated in Corollary (2.10). ☐

This simplification actually weakens the bound since q^{r-k} is a weak upper bound on $\binom{q-r}{r-k}$ but can make it easier to work with the expressions.

2.5 Bounds on $Q_h^{(r)}(c)$

Now we obtain upper and lower bounds on $Q_h^{(r)}(c)$. These bounds can be directly obtained from the bounds on $C_h^{(r)}(q)$.

Theorem 2.11. *Let h be an (n,m)-hash function with $n \geq r$ and $m \geq 2$. Let $\tau = s_h^{(r)}(\mathsf{qmax}_h^{(r)})$. Let c be a real number such that $0 \leq c < 1$. Then*

$$c^{1/r}\left(\frac{r}{e}\right) m^{(\frac{r-1}{r})\mu_r(h)} \leq Q_h^{(r)}(c) \leq \left(\frac{2c}{\tau}\right)^{1/r} \cdot rm^{(\frac{r-1}{r})\mu_r(h)}, \tag{30}$$

the upper bound being true when $c < \mathsf{cmax}_h^{(r)}$.

Table 2: h is a hash function with $m = 2^{80}$ and $\mu_r(h) = 0.9$.

r	$\mathsf{cmax}_h^{(r)}$
2	5.67003×10^{-2}
3	2.93125×10^{-3}
4	8.69089×10^{-5}
5	1.67242×10^{-6}
6	2.25836×10^{-8}
7	2.25855×10^{-10}
8	1.74035×10^{-12}

Proof. From Theorem 2.9 we have

$$C_h^{(r)}(q) \leq \underbrace{\binom{q}{r} m^{-(r-1)\mu_r(h)}}_{U_h^{(r)}(q)}.$$

To get the lower bound of Equation (30) we need to solve for q in the equation $U_h^{(r)}(q) = c$.

$$c = \binom{q}{r} m^{-(r-1)\mu_r(h)} \leq \left(\frac{qe}{r}\right)^r \frac{1}{m^{(r-1)\mu_r(h)}}$$

and so $q \geq c^{1/r} \left(\frac{r}{e}\right) m^{\left(\frac{r-1}{r}\right)\mu_r(h)}$. This proves the lower bound of Equation (30).

Similarly the upper bound can be obtained by finding the minimum value of q such that $L_h^{(r)}(q) \geq c$. Since the maximum value of $L_h^{(r)}(q)$ is $\mathsf{cmax}_h^{(r)}$ the minimum such q will be less than $\mathsf{qmax}_h^{(r)}$. By definition, $s_h^{(r)}(q)$ is decreasing in q which implies $s_h^{(r)}(q) > \tau$ for $q < \mathsf{qmax}_h^{(r)}$. Combining this with Theorem 2.9 we have for $q \geq \mathsf{qmax}_h^{(r)}$,

$$C_h^{(r)}(q) \geq \max_{r \leq t \leq q} L_h^{(r)}(t) \geq \frac{1}{2} s_h^{(r)}(q) \cdot \binom{q}{r} \cdot p_r \geq \frac{1}{2} s_h^{(r)}(q) \cdot \left(\frac{q}{r}\right)^r p_r \geq \frac{\tau}{2} \cdot \left(\frac{q}{r}\right)^r p_r.$$

If q is such that $C_h^{(r)}(q) \geq (\tau/2)(q/r)^r p_r \geq c$, then $Q_h^{(r)}(c) \leq q$. Let the minimum such q be denoted q^*. Clearly q^* is a solution to $(\tau/2)(q/r)^r p_r = c$. The upper bound on $Q_h^{(r)}(c)$ follows from this. $\square$

Theorem 2.11 establishes our claim that the number of trials required to find r-collisions with a significant probability of success is $\Theta\left(rm^{\left(\frac{r-1}{r}\right)\mu_r(h)}\right)$. For a given hash function h, the number of trials required to obtain an r-collision with a given probability c is at least as much as the lower bound. Also for $c \leq \mathsf{cmax}_h^{(r)}$, the number of trials required to obtain success probability c will not exceed the upper bound on $Q_h^{(r)}(c)$. For values of c greater than $\mathsf{cmax}_h^{(r)}$ we are unable to say anything about the maximum number of trials required to attain success probability c. But, the lower bound still continues to hold, i.e., we are still able to say that at least those many queries will be required to attain success probability c.

It would be interesting to know the range of values of c for which the upper bound on $Q_h^{(r)}(c)$ holds for different values of r. Because of the form of $L_h^{(r)}(q)$, we are unable to get a closed form expression for

Table 3: $n = 2^{512}$, $m = 2^{160}$ and $c = 0.78$.

$\mu_4(h)$	Lower bound on $Q_h^{(4)}(c)$		
0.22	1.22456×10^8	$\approx$	2^{26}
0.33	1.15233×10^{12}	$\approx$	2^{40}
0.44	1.08436×10^{16}	$\approx$	2^{53}
0.55	1.0204×10^{20}	$\approx$	2^{66}
0.66	9.60207×10^{23}	$\approx$	2^{79}
0.77	9.03568×10^{27}	$\approx$	2^{92}
0.88	8.5027×10^{31}	$\approx$	2^{106}
0.99	8.00115×10^{35}	$\approx$	2^{119}

$\mathsf{cmax}_h^{(r)}$. Table 2 shows how $\mathsf{cmax}_h^{(r)}$ varies with r when m and $\mu_r(h)$ are fixed. One can observe that the value of $\mathsf{cmax}_h^{(r)}$ is decreasing rapidly with increasing values of r which means that as r grows larger the upper bound of Theorem 2.11 is valid across smaller ranges of c.

Sensitivity of $Q_h^{(r)}(c)$ to r-balance. We provide some computational results that indicate how the number of trials required by the generic multi-collision attack changes according to the r-balance of the function being attacked. Table 3 shows the lower bound on $Q_h^{(4)}(c)$ for a fixed c and for functions with different values of 4-balance. The table indicates that for functions with higher 4-balance it is harder to find 4-collisions using the generic multi-collision attack when compared to functions with low 4-balance.

2.6 Applicability of our Results

It is practically infeasible to compute or estimate the r-balance of a given hash function. In [BK04], the authors address this question for the case $r = 2$. Some experiments are performed on SHA-1 and the balance is computed considering only small blocks of the output string. The details are as follows. Let $\mathsf{SHA}_n : \{0,1\}^n \rightarrow \{0,1\}^{160}$ denote the restriction of SHA-1 to inputs of length $n < 264$. Let $\mathsf{SHA}_{n;t_1 \cdots t_2} : \{0,1\}^n \rightarrow \{0,1\}^{t_2 t_1 + 1}$ denote the function that returns the t_1-th through t_2-th output bits of SHA_n . Bellare and Kohno ask what exactly is the balance of $\mathsf{SHA}_{32;t_1 \cdots t_2}$ when $t_2 - t_1 + 1 \in \{8, 16, 24\}$ and whether the functions $\mathsf{SHA}_{m;t_1 \cdots t_2}, m \in \{160, 256, 1024, 2048\}$, appear regular when $t_2 - t_1 + 1 \in \{8, 16, 24\}$. They calculate the balance of $\mathsf{SHA}_{32;t_1 \cdots t_2}$ for all pairs t_1, t_2 such that $t_2 - t_1 + 1 \in \{8, 16, 24\}$ and t_1 begins on a byte boundary (i.e., they look at all 1-, 2-, and 3-byte portions of the SHA-1 output). The values they calculate indicate that, for the specified values of t_1, t_2 , the balance of $\mathsf{SHA}_{32;t_1 \cdots t_2}$ is high.

However, these experiments do not provide any information about the balance of the actual SHA-1. Instead of pondering over how to compute the r-balance of a given function, we discuss what could be done to ensure a hash function has certain r-balance.

Proposition 2.12. *A hash function will have r-balance at least ν, if it is constructed in such a way that no range point has more than $n/m^{(\nu(r-1)+1)/r}$ pre-images.*

Proof. Essentially, we want the following to hold:

$$\mu_r = \frac{1}{r-1} \log_m \frac{n^r}{\sum_{i=1}^{m} (n_i)_r} \geq \nu$$

or equivalently, $\sum_{i=1}^{m}(n_i)_r \leq n^r m^{-(r-1)\nu}$. Suppose that for all $i \in \{1, 2, \cdots, m\}$, $n_i \leq n/m^\delta$ for some $\delta \in [0, 1]$. Then

$$\sum_{i=1}^{m}(n_i)_r \leq m\left(\frac{n}{m^\delta}\right)_r \leq m\left(\frac{n}{m^\delta}\right)^r.$$

To ensure r-balance at least ν, it is enough if δ is such that $m\left(\frac{n}{m^\delta}\right)^r \leq n^r m^{-(r-1)\nu}$. Solving for δ, we obtain, $\delta \geq (1/r) \times (\nu(r-1)+1)$. This completes the proof. □

So while building a hash function with r-balance at least ν, the designer must ensure that no range point has more than $n/m^{(\nu(r-1)+1)/r}$ pre-images. According to Theorem 2.11, this makes sure that an attacker will need at least $c^{1/r}(r/e)m^{\nu(r-1)/r}$ trials to obtain an r-collision with probability at least c. From Proposition 2.4, ensuring a lower bound on r_2-balance also provides a lower bound on r_1-balance. So, a designer can pick a suitable r_2 and ensure that the hash function has a high enough r_2-balance.

For example, suppose we want a 256-bit hash function with 180-bit security against the generic multi-collision attack for 4-collisions. In other words, we would like the attack to make at least 2^{180} queries to ensure that a 4-collision is obtained with a constant probability of success c. So, we require $c^{1/r}(r/e)m^{\nu(r-1)/r} > 2^{180}$ where $m = 2^{256}$ and $r = 4$. Taking logarithms to base two and assuming $c \approx 1$, under reasonable approximations, the 4-balance μ_4 must be at least $(4/3) \times 180/256 = 0.9375$. We can ensure this by designing the function in such a way that every range point has at least $n \times 2^{-244}$ pre-images. Clearly, this method can be extended for higher values of r.

3 Random Functions

In this section, we consider a uniform random (n, m) function. Our purpose is two-fold.

1. First, we would like to address the question of whether the balance of most functions is close to one. A combinatorial approach to this problem would be to count the number of functions which have balance close to one and then estimate this count as a fraction of the total number of (n, m) functions. Such a direct counting method is extremely difficult to carry out. Instead we adopt a probabilistic approach. The balance of a random function is a random variable. We can then study its expectation, variance and more importantly probability concentration bounds using Chebyshev's inequality and Chernoff bound. This analysis shows that the probability that the balance of a random function is close to one is itself very close to one. From this we conclude that most functions have balance close to one.

2. The second reason for considering random functions is to study the efficacy of the generic multi-collision attack in Section 2 on random functions. For the case of $r = 2$, this was investigated by Bellare and Kohno [BK04]. We generalize their result to the case of $r > 2$ and for $r = 2$ provide somewhat improved result.

3.1 Distribution of r-Balance

Consider a hash function picked uniformly at random from the set of all (n, m)-functions. A natural question that arises - how close is the r-balance of this function to 1? Also, it would be of general interest to know how r-balance is distributed for uniform random functions.

Consider the experiment of picking an (n, m)-function uniformly at random from the set of all (n, m) functions. A more convenient way to view this is the following: for each domain point, pick a range point independently and uniformly at random. Now, the number of pre-images of any range point is a random variable. By N_i, $i = 1, \ldots, m$, we denote the number of pre-images of the i-th range point. For a fixed i, define indicator random variables U_j for $j \in \{1, 2, \cdots, n\}$ as follows: $U_j = 1$ if x_j maps to y_i and 0 otherwise. Clearly $N_i = \sum_{j=1}^{n} U_j$. Since U_j is a Bernoulli trial with success probability $1/m$, N_i is binomially distributed, i.e., $N_i \sim \text{Bin}\left(n, \frac{1}{m}\right)$ having probability density function given by

$$\Pr[N_i = k] = \binom{n}{k} \frac{1}{m^k} \left(1 - \frac{1}{m}\right)^{n-k}.$$

Note that the N_i's are identically distributed but not independent due to the constraint that $\sum_{i=1}^{m} N_i = n$.

For any positive integer r, define $P_r = (\sum_{i=1}^{m} (N_i)_r)/n^r$ and $\Lambda_r = 1/(r-1) \times \log_m(1/P_r)$. Then P_r and Λ_r are random variables defined from the random variables N_i's. This Λ_r is the balance of a uniform random function. We are interested in finding the probability of Λ_r being close to 1. For this, we only need to look at the distribution of the N_i's and their falling factorials. First, we shall prove some identities involving the falling factorials of a random variable which follows the binomial distribution. Somewhat surprisingly, we were not able to locate these results in the literature.

Proposition 3.1. *Let $X, Y \sim \text{Bin}(n, \xi)$. Then the following statements are true.*

1. $\text{E}[(X)_r] = (n)_r\, \xi^r$.
2. $(n)_r^2 \xi^{2r} \leq \text{E}[{(X)_r}^2] < (n)_r\, \xi^r r!\, (n\xi + 1 - \xi)^r$.
3. $\text{E}[(X)_r (Y)_r] = (n)_{2r}\, \xi^{2r}$.

Proof. We will now prove each of the results stated above.

1. $\text{E}[(X)_r] = \sum_{k=0}^{n} (k)_r \binom{n}{k} \xi^k (1-\xi)^{n-k} = \xi^r \sum_{k=r}^{n} (k)_r \binom{n}{k} \xi^{k-r} (1-\xi)^{n-k}$. Differentiating both sides of $(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}$ r times partially with respect to x gives us $(n)_r (x+y)^{n-r} = \sum_{k=r}^{n} (k)_r \binom{n}{k} x^{k-r} y^{n-k}$. Substituting $x = \xi$ and $y = 1 - \xi$, we obtain $\text{E}[(X)_r] = (n)_r\, \xi^r$.

2. $$\text{E}[{(X)_r}^2] = \sum_{k=0}^{n} (k)_r (k)_r \binom{n}{k} \xi^k (1-\xi)^{n-k} = \xi^r \sum_{k=r}^{n} (k)_r (k)_r \binom{n}{k} \xi^{k-r} (1-\xi)^{n-k}.$$

   Multiplying both sides of the identity $(n)_r (x+y)^{n-r} = \sum_{k=r}^{n} (k)_r \binom{n}{k} x^{k-r} y^{n-k}$ by x^r, we get $(n)_r x^r (x+y)^{n-r} = \sum_{k=r}^{n} (k)_r \binom{n}{k} x^k y^{n-k}$. Let $g_0(x) = x^r (x+y)^{n-r}$ and define $g_s(x) = \frac{\partial g_{s-1}}{\partial x}$. Clearly,

   $$(n)_r g_r(x) = \sum_{k=r}^{n} (k)_r (k)_r \binom{n}{k} x^{k-r} y^{n-k}. \tag{31}$$

   It can be shown by induction that

   $$g_s(x) = (x+y)^{n-r-s} \sum_{j=0}^{s} \binom{s}{j} (n-j)_{s-j} (r)_j x^{r-j} y^j. \tag{32}$$

Combining Equations (31) and (32) and substituting $x = \xi$ and $y = 1 - \xi$, we get

$$\begin{aligned}
\mathrm{E}[(X)_r{}^2] &= (n)_r\, \xi^r \sum_{j=0}^{r} \binom{r}{j} (n-j)_{r-j} (r)_j \xi^{r-j} (1-\xi)^j \\
&< (n)_r\, \xi^r r! \sum_{j=0}^{r} \binom{r}{j} n^{r-j} \xi^{r-j} (1-\xi)^j \\
&= (n)_r\, \xi^r r!\, (n\xi + 1 - \xi)^r
\end{aligned}$$

The lower bound follows from Jensen's inequality.

3.

$$\begin{aligned}
\mathrm{E}[(X)_r (Y)_r] &= \sum_{0 \le k,l \le n} (k)_r (l)_r \frac{n!}{k!\, l!\, (n-l-k)!} \xi^k \xi^l (1-2\xi)^{n-k-l} \\
&= \xi^{2r} \sum_{r \le k,l \le n} (k)_r (l)_r \frac{n!}{k!\, l!\, (n-l-k)!} \xi^{k-r} \xi^{l-r} (1-2\xi)^{n-k-l}.
\end{aligned}$$

Consider the trinomial expansion of $(x+y+z)^n$

$$(x+y+z)^n = \sum_{0 \le k,l \le n} \frac{n!}{k!\, l!\, (n-k-l)!} x^k y^l z^{n-k-l}. \tag{33}$$

Differentiating partially (33) r times with respect to x gives us

$$(n)_r (x+y+z)^{n-r} = \sum_{\substack{0 \le l \le n \\ r \le k \le n}} (k)_r \frac{n!}{k!\, l!\, (n-k-l)!} x^{k-r} y^l z^{n-k-l}.$$

Differentiating partially the above equation r times with respect to y gives us

$$(n)_{2r} (x+y+z)^{n-2r} = \sum_{r \le k,l \le n} (k)_r (l)_r \frac{n!}{k!\, l!\, (n-k-l)!} x^{k-r} y^{l-r} z^{n-k-l}.$$

Substituting $x = \xi$, $y = \xi$ and $z = 1 - 2\xi$, we obtain $\mathrm{E}[(X)_r (Y)_r] = (n)_{2r}\, \xi^{2r}$.

□

We note that it is possible to obtain a closed form expression for $E[(X)_r^2]$ in terms of the hypergeometric function. But, this does not seem to be useful in the present context and so, we did not pursue this further. Using Proposition 3.1(1), we get the expected value of P_r as follows.

$$\mathrm{E}[P_r] = \frac{1}{n^r} \mathrm{E}\left[\sum_{i=1}^{m} (N_i)_r\right] = \frac{1}{n^r} \sum_{i=1}^{m} \mathrm{E}[(N_i)_r] = \frac{m}{n^r} \frac{(n)_r}{m^r} = \frac{(n)_r}{n^r} \frac{1}{m^{r-1}}. \tag{34}$$

From Proposition 3.1 we can obtain an upper bound on the variance P_r as follows.

$$
\begin{aligned}
\mathrm{Var}[P_r] &= \mathrm{Var}\left[\frac{\sum_{i=1}^m (N_i)_r}{n^r}\right] \\
&= \frac{1}{n^{2r}} \mathrm{E}\left[\left(\sum_{i=1}^m (N_i)_r\right)^2\right] - (\mathrm{E}[P_r])^2 \\
&= \frac{1}{n^{2r}} \mathrm{E}\left[\sum_{i=1}^m (N_i)_r{}^2 + \sum_{i\neq j} (N_i)_r (N_j)_r\right] - \frac{(n)_r{}^2}{n^{2r} m^{2r-2}} \\
&= \frac{1}{n^{2r}} \left(m\mathrm{E}[(N_1)_r{}^2] + m(m-1)\mathrm{E}[(N_1)_r (N_2)_r]\right) - \frac{(n)_r{}^2}{n^{2r} m^{2r-2}} \\
&< \frac{1}{n^{2r}} \left(m \frac{(n)_r}{m^{2r}} (n+m-1)^r + m(m-1) \frac{(n)_{2r}}{m^{2r}}\right) - \frac{(n)_r{}^2}{n^{2r} m^{2r-2}} \\
&= \frac{(n)_r}{n^{2r} m^{2r-2}} \left(\frac{(n+m-1)^r}{m} + (n-r)_r \left(1 - \frac{1}{m}\right) - (n)_r\right) \\
&< \frac{(n)_r (n+m-1)^r}{n^{2r} m^{2r-1}}
\end{aligned}
$$

Now consider the random variable Λ_r. We seek an upper bound on $\Pr[\Lambda_r < 1 - \varepsilon]$ where $0 < \varepsilon < 1$. The first simple result is the following.

Proposition 3.2. *For* $0 < \varepsilon < 1$, $\quad \Pr[\Lambda_r < 1 - \varepsilon] \leq \frac{(n)_r}{n^r} \frac{1}{m^{\varepsilon(r-1)}}$.

Proof. Using the definition of balance, we get

$$\Pr[\Lambda_r < 1 - \varepsilon] = \Pr[m^{-(r-1)\Lambda_r} > m^{-(r-1)(1-\varepsilon)}] = \Pr[P_r > m^{-(r-1)(1-\varepsilon)}]$$

By Markov's inequality, we have $\Pr[P_r > m^{-(r-1)(1-\varepsilon)}] \leq \dfrac{\mathrm{E}[P_r]}{m^{-(r-1)(1-\varepsilon)}} = \dfrac{(n)_r}{n^r} \dfrac{1}{m^{\varepsilon(r-1)}}$. ☐

For practical hash functions $(n)_r/n^r$ is almost 1. Suppose that $m = 2^{160}$, $r = 4$ and $\varepsilon = 0.01$, then we have $\Pr[\Lambda_4 < 0.99] \leq 1/2^{4.8} < 0.035$ which is quite low. This suggests that for most functions the 4-balance may be close to 1. But we need a stronger statement to substantiate this claim.

Proposition 3.3. *For* $0 < \varepsilon < 1$, $\quad \Pr[\Lambda_r < 1 - \varepsilon] < \frac{(n)_r}{n^r m^{2\varepsilon(r-1)}} e^{\frac{r(m-1)}{n}}$.

Proof. An application of Chebyshev's inequality gives us the following.

$$\begin{aligned}
\Pr\left[P_r > m^{-(r-1)(1-\varepsilon)}\right] &< \frac{\mathrm{E}[{P_r}^2]}{m^{-2(r-1)(1-\varepsilon)}} \\
&= \frac{m^{2(r-1)(1-\varepsilon)}}{n^{2r}} \left(m\mathrm{E}[{(N_1)_r}^2] + m(m-1)\mathrm{E}[(N_1)_r(N_2)_r]\right) \\
&= \frac{m^{2(r-1)(1-\varepsilon)}}{n^{2r}} \left(m\frac{(n)_r(n+m-1)^r}{m^{2r}} + (m^2-m)\frac{(n)_{2r}}{m^{2r}}\right) \\
&= \frac{(n)_r}{n^{2r}m^{2\varepsilon(r-1)}} \left(\frac{(n+m-1)^r}{m} + (n-r)_r\left(1-\frac{1}{m}\right)\right) \\
&< \frac{(n)_r}{n^{2r}m^{2\varepsilon(r-1)}} \left(\frac{(n+m-1)^r}{m} + (n+m-1)^r\left(1-\frac{1}{m}\right)\right) \\
&= \frac{(n)_r(n+m-1)^r}{n^{2r}m^{2\varepsilon(r-1)}} \\
&= \frac{(n)_r}{n^r m^{2\varepsilon(r-1)}} \left(1+\frac{m-1}{n}\right)^r \\
&< \frac{(n)_r}{n^r m^{2\varepsilon(r-1)}} e^{\frac{r(m-1)}{n}}. \qquad (35)
\end{aligned}$$

□

Observe that the bound of Proposition 3.3 is almost a square of the bound given by Proposition 3.2 which makes it better. Consider, for example, $n = 2^{512}$, $m = 2^{160}$, $r = 4$ and $\varepsilon = 0.01$. As mentioned earlier, for practical hash functions $(n)_r/n^r$ is almost 1. Ignoring this term, we get $\Pr[\Lambda_4 < 0.99] \leq 0.0035$ which is better compared to the Markov bound. As $\varepsilon \to 1$ the bound on $\Pr[\Lambda_r < 1-\varepsilon]$ decreases exponentially and faster than the bound of Proposition 3.2. This reinforces our claim that most functions have r-balance near 1.

For a stronger statement, we will use the results from Section 2.6 and perform a Chernoff bound based analysis. In the earlier analysis we showed that $\Pr[\Lambda_r < 1-\varepsilon]$ is small. Here we will show that $\Pr[\Lambda_r \geq 1-\varepsilon]$ is large.

Proposition 3.4. *For* $0 \leq \varepsilon \leq 1$,

$$\Pr[\Lambda_r \geq 1-\varepsilon] \geq 1 - \exp\left(\ln m + \frac{n}{m^{1-\varepsilon\left(\frac{r-1}{r}\right)}}\left(1-\varepsilon\left(\frac{r-1}{r}\right)\ln m\right) - \frac{n}{m}\right). \qquad (36)$$

Proof. Let $\delta = 1-\varepsilon(r-1)/r$. From Proposition 2.12, it follows that if $N_i \leq n/m^\delta$ for all $i \in \{1, \cdots, m\}$ then $\Lambda_r \geq 1-\varepsilon$. As a result we have,

$$\Pr[\Lambda_r \geq 1-\varepsilon] \geq \Pr\left[\bigwedge_{i=1}^{m}\left(N_i \leq \frac{n}{m^\delta}\right)\right] = 1 - \Pr\left[\bigvee_{i=1}^{m}\left(N_i > \frac{n}{m^\delta}\right)\right] \geq 1 - m\cdot\Pr\left[N_1 > \frac{n}{m^\delta}\right].$$

We now need to show that $m\Pr[N_1 > n/m^\delta]$ is "vanishingly" small. Let $\Delta = m^{1-\delta}-1$. Then $\Pr[N_1 > \frac{n}{m^\delta}] = \Pr[N_1 > (1+\Delta)\frac{n}{m}]$. We know that $\mathrm{E}[N_1] = n/m$.

Using the standard form of Chernoff bound (more details can be found in [MR95, pp. 68], we get

$$\begin{aligned} m\Pr[N_1 > (1+\Delta)\mathrm{E}[N_1]] &< m\left(\frac{e^{\Delta}}{(1+\Delta)^{1+\Delta}}\right)^{\mathrm{E}[N_1]} \\ &= m\frac{1}{e^{n/m}}\left(\frac{e}{1+\Delta}\right)^{(1+\Delta)n/m} \\ &= \exp\left(\ln m + (1+\Delta)\frac{n}{m}\left(1-\ln(1+\Delta)\right) - \frac{n}{m}\right) \\ &= \exp\left(\ln m + \frac{n}{m^{\delta}}\left(1-(1-\delta)\ln m\right) - \frac{n}{m}\right) \\ &= \exp\left(\ln m + \frac{n}{m^{1-\varepsilon\left(\frac{r-1}{r}\right)}}\left(1-\varepsilon\left(\frac{r-1}{r}\right)\ln m\right) - \frac{n}{m}\right). \end{aligned}$$

This completes the proof. □

For values of $\varepsilon \geq \frac{r}{(r-1)\ln m}$, the second term in the exponent is negative and hence $m\Pr[N_1 > n/m^{\delta}]$ decreases exponentially. This implies that $\Pr[\Lambda_r \geq 1-\varepsilon]$ will be close to 1 indicating that most functions have balance close to 1. Consider, for example, $\varepsilon = \frac{r}{(r-1)\ln m}$. Then Proposition 3.4 shows that $\Pr[\Lambda_r \geq 1 - r/((r-1)\ln m)] \geq 1 - m/\exp(n/m)$. The quantity $m/\exp(n/m)$ is vanishingly small, so that with overwhelming probability Λ_r is greater than $1 - r/((r-1)\ln m)$. Taking concrete values, let $m = 2^{160}$, $n = 2^{512}$ and $r = 4$. Then the probability that Λ_4 is greater than 0.99 is at least $1 - \exp(111 - 2^{352})$.

Even as ε becomes less than $\frac{r}{(r-1)\ln m}$ the probability will be very close to one as long as the expression within the exponent of (36) remains negative. At the point where the value within the exponent changes sign, the lower bound on probability becomes a huge negative quantity. We find this sudden change of the bound from being very close to one to a huge negative quantity to be a surprising feature. But, there does not seem to be any easy way to determine this "knee" point. On the other hand, the range of values for which this bound is valid seems to substantiate that the r-balance of most functions is concentrated near one.

3.2 Generic Attack on Random Functions

Consider a uniform random (n, m)-hash function. We consider the resistance of such a hash function to the generic multi-collision attack mentioned in Section 2. Our aim is to show that the attack works better against uniform random functions compared to regular functions. This is shown by proving that the success probability of the attack is higher for a uniform random function than for a regular function. For $r = 2$, this was shown by Bellare and Kohno. Informally, one may consider that having higher success probability means that it is easier to find r-collisions.

Let $C^{\$(r)}_{n,m}(q)$ be the probability that the generic multi-collision attack on a uniform random (n, m)-hash function succeeds in q trials. Here the probability is over the choice of the function and the points picked by the attack. Similarly, let $Q^{\$(r)}_{n,m}(c)$ denote the minimum number of trials required to obtain an r-collision with probability greater than or equal to c.

Let $h : X \to Y$ be a uniform random function i.e., for any $x \in X$ and $y \in Y$, $\Pr[h(x) = y] = 1/m$. Consider the experiment of choosing r elements independently and uniformly at random from the domain X. Let $p^{\$}_r$ denote the probability that these r elements form an r-collision. Let r elements $w_1, w_2, \cdots, w_r$ be picked independently and uniformly at random from the domain X. If A is the event that these are distinct and B is the event that $h(w_1) = \cdots = h(w_r)$, then $p^{\$}_r = \Pr[A] \cdot \Pr[B]$. Clearly,

$$\Pr[A] = \frac{(n)_r}{n^r} \text{ and } \Pr[B] = m \cdot \frac{1}{m^r}$$

since there are m choices for the common image. Thus we have,

$$p_r^{\$} = \frac{(n)_r}{n^r} \cdot \frac{1}{m^{r-1}}.$$

Note that from (34), $p_r^{\$}$ is equal to the expectation of P_r. This is not surprising, since both these quantities relate to the probability of an r-collision for a uniform random (n, m) function, although in different ways.

Consider the generic multi-collision attack on a uniform random (n, m)-hash function. The bounds on $C_{n,m}^{\$(r)}(q)$ and $Q_{n,m}^{\$(r)}(c)$ are obtained in a manner similar to that for a concrete hash function and we state some of the results without proofs.

Lemma 3.5. *Let ℓ be an integer such that $\ell > r$. Then $p_\ell^{\$} \leq (p_r^{\$})^{\ell/r}$*

Theorem 3.6. *For a uniform random (n, m)-hash function with $n > r$ the following holds.*

$$\max_{r \leq t \leq q} L_{n,m}^{\$(r)}(t) \leq C_{n,m}^{\$(r)}(q) \leq \binom{q}{r} \cdot p_r^{\$} \tag{37}$$

where the function $L_{n,m}^{\$(r)}(t)$ is defined as follows:

$$L_{n,m}^{\$(r)}(t) = \frac{1}{2}\left(2 - \sum_{k=0}^{r-1} \binom{r}{k}\binom{t-r}{r-k}(p_r^{\$})^{(r-k)/r}\right) \cdot \binom{t}{r} \cdot p_r^{\$} \tag{38}$$

For the purpose of comparison to regular functions we will use a simplified version of the lower bound on $C_{n,m}^{\$(r)}(q)$. This is obtained in a manner similar to the one given in the proof of Corollary 2.10.

Corollary 3.7. *For a uniform random (n, m)-hash function with $n > r$, let*

$$\alpha^{\$}(q) = qm^{-\left(\frac{r-1}{r}\right)}. \tag{39}$$

Then

$$C_{n,m}^{\$(r)}(q) \geq \max_{r \leq t \leq q} \frac{1}{2}\left(3 - (\alpha^{\$}(t) + 1)^r\right) \cdot \binom{t}{r} \cdot p_r^{\$} \tag{40}$$

We now proceed towards obtaining bounds on $Q_{n,m}^{\$(r)}(c)$. The upper bound can be obtained the same way as in Theorem 2.11. Only a proof of the lower bound is provided here.

Theorem 3.8. *Consider a uniform random (n, m)-hash function with $n > r$ and let c be a real number such that $0 \leq c < 1$. Then*

$$c^{1/r} r \cdot e^{\left(\frac{r-1}{2n} - 1\right)} \cdot m^{\left(\frac{r-1}{r}\right)} \leq Q_{n,m}^{\$(r)}(c) \leq \min\{q : L_{n,m}^{\$(r)}(q) = c\}, \tag{41}$$

the upper bound being true when

$$c < \mathsf{cmax}_r^{\$}(n, m). \tag{42}$$

where $\mathsf{cmax}_r^{\$}(n, m)$ denotes the maximum positive value that the function $L_{n,m}^{\$(r)}(q)$ attains.

Proof. From Theorem 3.6 we have

$$C_{n,m}^{\$(r)}(q) \le \underbrace{\binom{q}{r} p_r^{\$}}_{U_{n,m}^{\$(r)}(q)}$$

To get the lower bound of Equation (41) we need to solve for q in the equation $U_{n,m}^{\$(r)}(q) = c$.

$$\begin{aligned} c &= \binom{q}{r} \cdot \left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right) \cdots \left(1 - \frac{r-1}{n}\right) \cdot \frac{1}{m^{r-1}} \\ &\le \left(\frac{qe}{r}\right)^r e^{-1/n} e^{-2/n} \cdots e^{-(r-1)/n} \cdot \frac{1}{m^{r-1}} \\ &= \left(\frac{qe}{r}\right)^r e^{-r(r-1)/2n} \cdot \frac{1}{m^{r-1}} \end{aligned}$$

Solving for q in the above inequality will give

$$q \ge c^{1/r} r \cdot e^{\left(\frac{r-1}{2n} - 1\right)} \cdot m^{\left(\frac{r-1}{r}\right)}$$

$\square$

Comparison with regular functions. Let $C_{n,m}^{\mathrm{reg}(r)}(q)$ denote the probability of success of the generic multi-collision attack on a regular (n, m)-hash function. Let the maximum value of r-balance be denoted $\mu_r^{\max}$ and let the value of p_r corresponding to $\mu_r^{\max}$ be denoted as p_r^{reg}. Since all regular functions have the same value for p_r, we have $C_{n,m}^{\mathrm{reg}(r)}(q) = C_h^{(r)}(q)$ for some function h with maximum balance.

Lemma 3.9. *Let n, m and r be integers such that $r \ge 2$ and $n \ge rm$. Then*

$$\frac{n(n-1)\cdots(n-(r-1))}{n(n-m)\cdots(n-(r-1)m)} > 1 + \frac{m-1}{n-m} \cdot \frac{r(r-1)}{2}$$

Proof. The condition $n \ge rm$ ensures that the denominator $n(n-m)\cdots(n-(r-1)m)$ is non-zero.

$$\begin{aligned} \frac{n(n-1)\cdots(n-(r-1))}{n(n-m)\cdots(n-(r-1)m)} &= \frac{n-1}{n-m} \cdot \frac{n-2}{n-2m} \cdots \frac{n-(r-1)}{n-(r-1)m} \\ &= \left(1 + \frac{m-1}{n-m}\right) \cdot \left(1 + \frac{2(m-1)}{n-2m}\right) \cdots \left(1 + \frac{(r-1)(m-1)}{n-(r-1)m}\right) \\ &> 1 + \frac{m-1}{n-m} + \frac{2(m-1)}{n-2m} + \cdots + \frac{(r-1)(m-1)}{n-(r-1)m} \\ &> 1 + \frac{m-1}{n-m} + \frac{2(m-1)}{n-m} + \cdots + \frac{(r-1)(m-1)}{n-m} \\ &= 1 + \frac{m-1}{n-m} \cdot \frac{r(r-1)}{2} \end{aligned}$$

$\square$

Theorem 3.10. *Let $r \ge 2$ and $n \ge rm$ and*

$$\beta = 1 + \frac{m-1}{n-m} \cdot \frac{r(r-1)}{2}$$

Then

$$C^{\$(r)}_{n,m}(q) > C^{\mathrm{reg}(r)}_{n,m}(q) \tag{43}$$

for all q such that $q \leq \left(\left(3 - \frac{2}{\beta}\right)^{1/r} - 1\right) m^{(r-1)/r}$.

Proof. From (39) recall that $\alpha^{\$}(q) = qm^{-\left(\frac{r-1}{r}\right)}$ and by the bound given on q, we have $1/\beta \leq (3 - (\alpha^{\$}(q) + 1)^r)/2$. This will be used in the computation below. Also Lemma 3.9 is used in the last but one step of the computation.

From Corollary 3.7 and Lemma 3.9, we have

$$\begin{aligned}
C^{\$(r)}_{n,m}(q) &\geq \max_{r \leq t \leq q} \frac{1}{2}(3 - (\alpha^{\$}(t) + 1)^r)\binom{t}{r} p^{\$}_r \\
&\geq \frac{1}{2}(3 - (\alpha^{\$}(q) + 1)^r)\binom{q}{r} p^{\$}_r \\
&\geq \frac{1}{\beta}\binom{q}{r}\frac{(n)_r}{n^r}\frac{1}{m^{r-1}} \\
&= \frac{1}{\beta}\binom{q}{r}\frac{(n)_r}{n^r}\frac{1}{m^{r-1}} \\
&= \frac{1}{\beta}\binom{q}{r}\frac{(n)_r}{n^r}\frac{1}{m^{r-1}}\frac{n(n-m)\cdots(n-(r-1)m)}{n(n-m)\cdots(n-(r-1)m)} \\
&= \frac{1}{\beta}\frac{(n)_r}{n(n-m)\cdots(n-(r-1)m)}\binom{q}{r}\frac{m\left(\frac{n}{m}\right)_r}{n^r} \\
&> \frac{1}{\beta}\left(1 + \frac{m-1}{n-m}\cdot\frac{r(r-1)}{2}\right)\binom{q}{r} p^{\mathrm{reg}}_r \\
&\geq C^{\mathrm{reg}(r)}_{n,m}(q)
\end{aligned}$$

□

Theorem 3.10 shows that for a certain range of q, it is easier to find r-collisions for random functions than for regular functions. So, random functions provide lesser security compared to regular functions. The value of β is greater than 1 and consequently, the value of $(3 - 2/\beta)$ is also greater than 1 so that the upper bound on q required in Theorem 3.10 is not vacuous. So, for this range of q, it is easier to find r-collisions for uniform random functions than for regular functions. A similar result has been obtained by Bellare and Kohno [BK04] for $r = 2$, but only when n *equals* $2m$ and $m \geq 5$. For these values of the parameters, choosing $q \leq 0.37m^{1/2}$ satisfies the condition of Theorem 3.10 while the range of q obtained in [BK04] is $q \leq 0.1m^{1/2}$. Further, Theorem 3.10 holds for $n \geq rm$ and hence, even for $r = 2$, it is more general than [BK04].

Note. We would like to emphasize that we have considered only *generic* multi-collision attacks. For attacks which "look into" the structure of a hash function, a regular function may become completely vulnerable. So, the comparative strengths of random versus regular functions discussed above must not be taken to mean that regular functions are better than random functions against all kinds of attacks.

4 Expected Number of Trials to Obtain an r-Collision

In this section, we get back to the study of a concrete hash function as opposed to a uniform random function. Recall that the determining factor for an r-collision is the probability n_i/n that a randomly chosen domain point maps to the i-th range point.

Consider the following strategy. Points are picked from the domain one by one, until an r-collision is obtained. Clearly, the number of domain points picked is a random variable. We are interested in the expectation of this random variable. This is worked out in Section 4.1.

Interestingly, there is a much earlier work by Klamkin and Newman [KN67] which considers a similar problem. We interpret and extend their work in terms of a concrete hash function in Section 4.2.

4.1 Case of Actual r-Collisions

Let h denote the given hash function. Suppose the domain points are chosen independently and uniformly at random and h is applied to each of them. The process is continued as long as necessary until an r-collision occurs. We would then like to know the expected number of trials $E_h^{(r)}$ to obtain an r-collision.

For the case of $r = 2$, this was analysed by Bellare and Kohno. Given a hash function h, they denoted by E_h the expected number of trials required to obtain a collision (i.e., $E_h = E_h^{(2)}$) and obtained bounds on E_h. These bounds are obtained from two facts of a more general nature. They show that if $q \geq 2$ is the number of trials then $q(1 - C_h(q-1)) \leq E_h \leq q/C_h(q)$. The arguments used to obtain these bounds also go through for general r.

Proposition 4.1. *For any $q \geq r$,*

$$q(1 - C_h^{(r)}(q-1)) \leq E_h^{(r)} \leq \frac{q}{C_h^{(r)}(q)}.$$

Proof. The ideas involved in the proof are from [BK04]. Let $D_h^{(r)}(q)$ be the probability that the first r-collision is found at trial number q. Then $\sum_{i \geq q} D_h^{(r)}(i)$ is the probability that the first r-collision is found after $(q-1)$ trials which is equal to the probability that the first $(q-1)$ trials do not provide an r-collision. So, $\sum_{i \geq q} D_h^{(r)}(i) = 1 - C_h^{(r)}(q-1)$. Then

$$E_h^{(r)} = \sum_{i \geq 1} i D_h^{(r)}(i) \geq q \sum_{i \geq q} D_h^{(r)}(i) = q(1 - C_h^{(r)}(q-1)).$$

Obtaining the upper bound is a little more involved. Consider the trials to be conducted in batches of q trials each, i.e., trials with $x_{q(i-1)+1}, \ldots, x_{qi}$ are conducted in batch number i. Let $X_i = 1$ if an r-collision is found in batch number i and 0 otherwise. Since the x_js are chosen independently and uniformly at random, the random variables $X_1, X_2, \ldots$ are mutually independent Bernoulli trials with $\Pr[X_i = 1] = C_h^{(r)}(q)$ for all $i \geq 1$. Let Y be a random variable whose value is i if $X_i = 1$ and $X_k = 0$ for $1 \leq k \leq i-1$. Then Y follows the geometric distribution. Denote $C_h^{(r)}(q)$ by ε and then the expected value of qY can be computed as

$$\begin{aligned} \mathrm{E}[qY] &= q\varepsilon + 2q(1-\varepsilon)\varepsilon + \cdots + iq(1-\varepsilon)^{i-1}\varepsilon + \cdots \\ &= q\varepsilon\left(\frac{1}{\varepsilon^2}\right) = \frac{q}{\varepsilon} = \frac{q}{C_h^{(r)}(q)}. \end{aligned}$$

The above process of batching ignores the possibility that an r-collision can occur between the trials of batch number i and the trials of the previous batches. So, batching can only increase the expected number of trials to find an r-collision and hence

$$E_h^{(r)} \leq \mathrm{E}[qY] = \frac{q}{C_h^{(r)}(q)}.$$

This completes the proof. $\square$

To obtain more meaningful bounds, we need to evaluate the bounds in Proposition 4.1 for some values of q. A good value of q is $\mathsf{qmax}_h^{(r)}$ which is the point where the function $L_h^{(r)}(q)$ attains its maximum, i.e., the value of q for which the lower bound on $C_h^{(r)}(q)$ attains the maximum value $\mathsf{cmax}_h^{(r)}$. This gives the following bounds.

$$\mathsf{qmax}_h^{(r)}\left(1 - C_h^{(r)}\left(\mathsf{qmax}_h^{(r)} - 1\right)\right) \leq E_h^{(r)} \leq \frac{\mathsf{qmax}_h^{(r)}}{\mathsf{cmax}_h^{(r)}}.$$

As noted in Section 2.4, it is difficult to obtain a closed form expression for $\mathsf{qmax}_h^{(r)}$, so it is still difficult to understand what the above bounds really mean. Further, these bounds are not in terms of the balance. To get them in terms of the balance, we have to evaluate the bounds for suitable values of q. In fact, we evaluate the lower and upper bounds in Proposition 4.1 for different values of q.

Let $Q = m^{((r-1)/r)\mu_r(h)}$. Then from Theorem 2.6

$$C_h^{(r)}(Q-1) \leq \binom{Q-1}{r} p_r \leq \frac{Q^r p_r}{r!} = \frac{1}{r!}.$$

This shows

$$E_h^{(r)} \quad \geq \quad \left(1 - \frac{1}{r!}\right) m^{\frac{(r-1)}{r}\mu_r(h)}. \tag{44}$$

The upper bound involves a little more calculation. From Corollary 2.10 we have that, for $\alpha(q) = qm^{-\left(\frac{r-1}{r}\right)\mu_r(h)}$,

$$C_h^{(r)}(q) \geq \frac{1}{2}\left(3 - (\alpha(q)+1)^r\right) \cdot \binom{q}{r} \cdot m^{-(r-1)\mu_r(h)}.$$

Put $\alpha(q) = \delta_r$. We specify the exact value of δ_r later.

For $q \geq r$, we have $(1 - (r-1)/q) \geq 1/r$ and so

$$\begin{aligned}
\frac{q!}{r!(q-r)!} = \frac{q(q-1)\cdots(q-r+1)}{r!} \quad &= \quad \frac{q^r}{r!}\left(1 - \frac{1}{q}\right)\cdots\left(1 - \frac{r-1}{q}\right) \\
&\geq \quad \frac{q^r}{r!}\left(1 - \frac{r-1}{q}\right)^{r-1} \geq \frac{q^r}{r!r^{r-1}}.
\end{aligned}$$

Putting $q = \delta_r m^{((r-1)/r)\mu_r(h)} = \delta_r Q$, we have

$$\begin{aligned}
C_h^{(r)}(\delta_r Q) \quad &\geq \quad \frac{1}{2}(3 - (\delta_r+1)^r)\frac{\delta_r^r Q^r m^{-(r-1)\mu_r(h)}}{r!r^{r-1}} \\
&= \quad \frac{1}{2}\frac{(3 - (\delta_r+1)^r)\delta_r^r}{r!r^{r-1}}.
\end{aligned}$$

Proposition 4.1 now shows that

$$E_h^{(r)} \leq \frac{\delta_r Q}{C_h^{(r)}(\delta_r Q)} \leq \frac{2r!r^{r-1}}{\delta_r^r(3-(\delta_r+1)^r)} \times \delta_r m^{\frac{(r-1)}{r}\mu_r(h)}$$
$$= \frac{2r!r^{r-1}}{\delta_r^{r-1}(3-(\delta_r+1)^r)} \times m^{\frac{(r-1)}{r}\mu_r(h)}.$$

The value of δ_r is chosen such that it maximizes $x^{r-1}(3-(x+1)^r)$. This in turn, minimizes the upper bound. Differentiating $x^{r-1}(3-(x+1)^r)$ with respect to x and setting to zero, we obtain $x^{r-1}((2r-1)x+r-1)-3(r-1)=0$. (The solution $x=0$ has been ruled out.) The polynomial $x^{r-1}((2r-1)x+r-1)-3(r-1)$ has exactly one sign change and by Descartes' rule of signs has exactly one positive real root. We let δ_r to be the value of this root. Combining the two bounds leads to the following result.

Theorem 4.2. *Let h be an (n,m) hash function and δ_r be the positive real root of the polynomial $x^{r-1}((2r-1)x+r-1)-3(r-1)$. Then*

$$\left(1-\frac{1}{r!}\right) m^{\frac{(r-1)}{r}\mu_r(h)} \leq E_h^{(r)} \leq \frac{2r!r^{r-1}}{\delta_r^{r-1}(3-(\delta_r+1)^r)} \times m^{\frac{(r-1)}{r}\mu_r(h)}.$$

For $r=2$, δ_2 is the positive real root of $3x^2+4x-2=0$ and so $\delta_2=(\sqrt{5}-2)/3$. Using this, we obtain

$$\frac{1}{2}\cdot m^{\mu_2(h)/2} \leq E_h^{(2)} \leq 56\cdot m^{\mu_2(h)/2}.$$

Recall that $m^{-\mu_2(h)} = m^{-\mu(h)} - 1/n$. This can be used to translate bounds obtained in terms of $\mu(h)$ into bounds in terms of $\mu_2(h)$. For the sake of comparison, we do this for the bounds on $E_h = E_h^{(2)}$ obtained in [BK04].

$$\frac{1}{2}\cdot\sqrt{\frac{n}{n+m^{\mu_2(h)}}} \times m^{\mu_2(h)/2} \leq E_h^{(2)} \leq 72\cdot\sqrt{\frac{n}{n+m^{\mu_2(h)}}} \times m^{\mu_2(h)/2}.$$

Clearly, the bound that we obtain is better.

4.2 Case of Possibly Trivial r-Collisions

As before, let h be the hash function under consideration. Consider the scenario in which one picks points from the domain using uniform random sampling with replacement, applies h to these points and simply considers the event that r of the image points are equal. In this setting, we do not put the restriction that the pre-images of the r equal image points are necessarily distinct. So, this event does not necessarily give us an r-collision. Recall that earlier we had called this event to be a possibly trivial r-collision.

In a different context, this problem was studied much earlier by Klamkin and Newman [KN67]. They consider the following problem. Suppose m alternatives are equally likely. Consider the experiment of choosing from these repeatedly with replacement until one of the alternatives has occurred r times. The problem is to find $E(m,r)$, the expected number of repetitions necessary for this success. They show that, for large m, $E(m,r)$ is approximately $(r!)^{1/r}\Gamma(1+1/r)m^{(r-1)/r}$.

The setting of Klamkin and Newman can be considered to be finding possibly trivial r-collisions in the following manner. Suppose that points from the domain of h are chosen using uniform random sampling with replacement and h is applied to them. This corresponds to choosing points from the range with replacement. If h is a regular function, i.e., the probabilities n_i/n are all equal, then each of the range points is equally likely and the process corresponds to choosing the m range points uniformly and with replacement.

To tackle an arbitrary hash function, we need to consider the case of non-uniform probabilities. It turns out that by using a notion similar to r-balance, the analysis of Klamkin and Newman goes through for unequal probabilities. In the context of concrete hash functions, the asymptotic analysis of Klamkin and Newman is somewhat less meaningful. Instead, we obtain a lower bound on the expected number of choices to obtain a possibly trivial collision. We note that the generalisation of Klamkin and Newman's result to the setting of arbitrary probability distribution is of some interest in its own right.

Let $p_i = n_i/n$ and $\tilde{p} = (p_1, p_2, \cdots, p_m)$. Let $E(m, r, \tilde{p})$ denote the expected number of points chosen before obtaining a possibly trivial collision. The approach we use is identical to that of [KN67].

Identify the variables x_i with the m alternatives of the experiment so that the terms of the expansion of $(x_1 + \cdots + x_m)^q$ can be thought of as outcomes when the experiment is carried out q times. Let T_r denote the truncating operation which when applied to a polynomial, or a power series, has the effect of removing any term which contain some variable x_i raised to a power $\geq r$. If A and B are polynomials in the variables $x_1, \ldots, x_m$, then it is easy to see that $T_r(A + B) = T_r(A) + T_r(B)$. Further, if the sets of variables on which A and B depend are disjoint, then $T_r(AB) = T_r(A)T_r(B)$.

The expression $T_r\{(x_1 + \cdots + x_m)^q\}$ represents all possible outcomes of q experiments such that no alternative has appeared r or more times. Thus,

$$T_r\{(x_1 + \cdots + x_m)^q\}\big|_{p_1, p_2, \cdots, p_m}$$

is equal to the failure probability after q trials. It is known that the expected number of trails is exactly equal to the sum of these failure probabilities (refer to [Fel08, pp.265–266]). So

$$E(m, r, \tilde{p}) = \sum_{q=0}^{\infty} T_r\{(x_1 + \cdots + x_m)^q\}\big|_{p_1, p_2, \cdots, p_m}$$

Define $F(t) = \displaystyle\sum_{q=0}^{\infty} T_r\{(x_1 + \cdots + x_m)^q\}\frac{t^q}{q!}$. We then have

$$\begin{aligned}
F(t) &= T_r\left\{\sum_{q=0}^{\infty}(x_1 + \cdots + x_m)^q \frac{t^q}{q!}\right\} \\
&= T_r\left\{e^{(x_1 + x_2 + \cdots + x_m)t}\right\} \\
&= T_r\left\{e^{x_1 t}\right\} T_r\left\{e^{x_2 t}\right\} \cdots T_r\left\{e^{x_m t}\right\} \\
&= S_r(x_1 t) S_r(x_2 t) \cdots S_r(x_m t),
\end{aligned}$$

where $S_r(x)$ is the r-th partial sum of e^x, i.e., $S_r(x) = \sum_{j<r} \frac{x^j}{j!}$. Using the formula $1 = \displaystyle\int_0^{\infty} \frac{t^q}{q!} e^{-t} \mathrm{d}t$,

$$\sum_{q=0}^{\infty} T_r\{(x_1 + \cdots + x_m)^q\} = \int_0^{\infty} F(t) e^{-t} \mathrm{d}t = \int_0^{\infty} S_r(x_1 t) S_r(x_2 t) \cdots S_r(x_m t) e^{-t} \mathrm{d}t.$$

Setting $x_i = p_i$ for all i,

$$E(m, r, \tilde{p}) = \int_0^{\infty} S_r(p_1 t) S_r(p_2 t) \cdots S_r(p_m t) e^{-t} \mathrm{d}t. \tag{45}$$

We can write

$$S_r(p_1t)S_r(p_2t)\cdots S_r(p_mt)e^{-t} = S_r(p_1t)S_r(p_2t)\cdots S_r(p_mt)e^{-(p_1+\cdots+p_m)t} = \prod_{i=1}^{m} S_r(p_it)e^{-p_it}.$$

At this point, we need to analyse the expression $S_r(x)e^{-x}$. Note that $S_r(x) = 1+\frac{x}{1}+\frac{x^2}{2!}+\cdots+\frac{x^{r-1}}{(r-1)!}$.

$$\begin{aligned}
e^x &= 1+\frac{x}{1}+\frac{x^2}{2!}+\cdots+\frac{x^{r-1}}{(r-1)!} \\
&\quad +\frac{x^r}{r!}\left(1+\frac{x}{r+1}+\frac{x^2}{(r+2)(r+1)}+\cdots+\frac{x^{r-1}}{(2r-1)_{r-1}}\right) \\
&\quad +\frac{x^{2r}}{(2r)!}\left(1+\frac{x}{2r+1}+\frac{x^2}{(2r+2)(2r+1)}+\cdots+\frac{x^{r-1}}{(3r-1)_{r-1}}\right) \\
&\quad +\cdots \\
&< S_r(x)\left(1+\frac{x^r}{r!}+\frac{x^{2r}}{(2r)!}+\frac{x^{3r}}{(3r)!}+\cdots\right) \\
&< S_r(x)\left(1+\frac{x^r}{r!}+\frac{1}{2!}\left(\frac{x^r}{(r)!}\right)^2+\frac{1}{3!}\left(\frac{x^r}{(r)!}\right)^3+\cdots\right) \\
&= S_r(x)\exp\left(\frac{x^r}{r!}\right).
\end{aligned}$$

This shows that $S_r(x)e^{-x} > \exp(-x^r/r!)$. So,

$$\prod_{i=1}^{m} S_r(p_it)e^{-p_it} > \prod_{i=1}^{m}\exp\left(-\frac{(p_it)^r}{r!}\right) = \exp\left(-\frac{1}{r!}\sum_{i=1}^{m}(p_it)^r\right) = \exp\left(-\frac{t^r}{r!}\omega_r\right)$$

where $\omega_r = \sum_{i=1}^{m}(p_i)^r$. We define a notion which is similar to r-balance in the following manner:

$$\eta_r(h) = \frac{1}{r-1}\log_m \frac{1}{\sum_{i=1}^{m} p_i^r} \tag{46}$$

so that $\omega_r = m^{-(r-1)\eta_r}$. Now using (45), we have

$$E(m,r,\tilde{p}) = \int_0^\infty S_r(p_1t)S_r(p_2t)\cdots S_r(p_mt)e^{-t}\mathrm{d}t > \int_0^\infty \exp\left(-\frac{t^r}{r!}\omega_r\right)\mathrm{d}t.$$

Let $z = t^r\omega_r/r!$. Then

$$\begin{aligned}
\int_0^\infty e^{-t^r\omega_r/r!} &= \int_0^\infty e^{-z}\frac{r!}{r\omega}\frac{\omega^{1-1/r}}{(r!z)^{1-1/r}}\mathrm{d}z \\
&= \left(\frac{r!}{\omega_r}\right)^{1/r}\frac{1}{r}\int_0^\infty e^{-z}z^{\frac{1}{r}-1}\mathrm{d}z \\
&= \left(\frac{r!}{\omega_r}\right)^{1/r}\frac{1}{r}\Gamma\left(\frac{1}{r}\right) \\
&= m^{\frac{r-1}{r}\eta_r}\,(r!)^{1/r}\,\Gamma(1+r).
\end{aligned}$$

We summarise this in the following result.

Theorem 4.3. *For a fixed r,* $\quad E(m, r, \tilde{p}) > m^{\frac{r-1}{r}\eta_r(h)} (r!)^{1/r} \Gamma(1+r)$.

From Theorem 4.2, the expected number of steps for finding actual r-collisions is about $m^{((r-1)/r)\mu_r(h)}$ while from Theorem 4.3, the expected number of steps for finding possibly trivial r-collisions is about $m^{((r-1)/r)\eta_r(h)}$. For small r, these two values are not very different. But, as r grows the difference between them also grows. In the context of finding r-collisions, Theorem 4.2 is more relevant.

5 Conclusion

We have introduced the notion of r-balance of a concrete hash function h. This notion is used to quantify the resistance of h to generic multi-collision attack. Bounds are obtained on the success probability of finding r-collisions using q trials. These are then translated into bounds on the number of trials required for a desired success probability. A similar analysis for uniform random function shows that such functions offer less resistance compared to regular functions. The work in this paper extended earlier work by Bellare and Kohno [BK04] for collisions, i.e., for $r = 2$ to any $r \geq 2$. To a certain extent, we complete the work started by them.

Going beyond the theme set out by Bellare and Kohno, we propose a new design criteria for hash functions based on the notion of r-balance. Also, using tools from probability theory, we study the nature of the variation of r-balance over the set of all functions.

Acknowledgement

We thank the anonymous reviewers of a previous version of this paper for their comments as well as for pointing out references [KN67] and [STKT06] to us.

References

[BK04] M. Bellare and T. Kohno. Hash function balance and its impact on birthday attacks. In C. Cachin and J. Camanisch, editors, *Advances in Cryptology - EUROCRYPT '04*, volume 3027 of *Lecture Notes in Computer Science*. Springer-Verlag, 2004.

[BPVY00] E. Brickell, D. Pointcheval, S. Vaudenay, and M. Yung. Design validation for discrete logarithm based signature schemes. In *PKC'2000*, volume 1751 of *Lecture Notes in Computer Science*, pages 276–292. Springer-Verlag, 2000.

[Fel08] W. Feller. *An introduction to probability theory and its applications*, volume I. Wiley India, 3 edition, 2008.

[GS94] M. Girault and J. Stern. On the length of cryptographic hash-values used in identification schemes. In *Advances in Cryptology - CRYPTO 1994*, volume 839 of *Lecture Notes in Computer Science*, pages 202–215. Springer-Verlag, 1994.

[HS06] J. J. Hoch and A. Shamir. Breaking the ICE - finding multicollisions in iterated concatenated and expanded (ICE) hash functions. In *Fast Software Encryption 2006*, volume 4047 of *Lecture Notes in Computer Science*, pages 179–194, Berlin, Germany, 2006. Springer-Verlag.

[JL09] A. Joux and S. Lucks. Improved generic algorithms for 3-collisions. In *Advances in Cryptology ASIACRYPT 2009*, volume 5912 of *Lecture Notes in Computer Science*, pages 347–363. Springer Berlin, 2009.

[Jou04] A. Joux. Multicollisions in iterated hash functions. application to cascaded constructions. In *Advances in Cryptology - CRYPTO 2004*, volume 3152 of *Lecture Notes in Computer Science*, pages 474–490, Berlin, Germany, 2004. Springer-Verlag.

[KN67] M. S. Klamkin and D. J. Newman. Extensions to the birthday surprise. *Journal of Combinatorial Theory*, 3:279–282, 1967.

[Lev81] B. Levin. A representation for multinomial cumulative distribution functions. *The Annals of Statistics*, 9(5):1123–1126, September 1981.

[McK66] E. H. McKinney. Generalized birthday problem. *The American Mathematical Monthly*, 73(4):385–387, April 1966.

[MR95] R. Motwani and P. Raghavan. *Randomized Algorithms*, chapter 4. Cambridge University Press, 1 edition, 1995.

[NS07] M. Nandi and D. R. Stinson. Multicollision attacks on some generalized sequential hash functions. *IEEE transactions on Information Theory*, 53(2):759–767, February 2007.

[Pre93] B. Preneel. *Analysis and Design of Cryptograhic Hash Functions*. PhD thesis, Katholieke Universiteit Leuven, Leuven, Belgium, 1993.

[RS96] R. Rivest and A. Shamir. PayWord and MicroMint - two simple micropayment schemes. *CryptoBytes*, 2(1):7–11, Spring 1996.

[STKT06] K. Suzuki, D. Tonien, K. Kurosawa, and K. Toyota. Birthday paradox for multicollisions. In *Information Security and Cryptology ICISC 2006*, volume 4296 of *Lecture Notes in Computer Science*, pages 29–40. Springer Berlin, 2006.