

大样本序列重叠碰撞统计检验方法

陈东昱^{1),2)} 范丽敏¹⁾ 陈 华¹⁾ 王 舰^{1),2)} 付一方^{1),2)}

¹⁾(中国科学院软件研究所可信计算与信息保障实验室 北京 100190)

²⁾(中国科学院大学 北京 100049)

摘 要 随机数发生器(RNG)产生的随机数质量直接关系到密码系统的安全性,而随机数统计假设检验是常用的随机数质量评测方法,在密码应用实践中发挥着重要作用.近年来,信息技术的发展极大提升了大数据处理能力,RNG的随机数统计检验也面临着更大数据量、更强检验能力和更快检验速度的迫切需求.本文针对现有检验方法在大数据量(GB级)检验适用性方面的问题,对大样本下的长模板重叠碰撞类检验进行了深入研究.首先,我们构造了碰撞对数统计量,并在此基础上通过理论推导给出了一种适用各种计数方式的新型碰撞统计模型.继而,我们得到了 n 比特长序列 m 比特模板的重叠碰撞对数的理论期望值,并基于该期望值通过统计分析,提出了两种序列重叠碰撞检验方法:(1)我们基于重叠碰撞数期望可作为碰撞出现概率上界的事实,利用极小非平凡碰撞期望拒绝较长碰撞出现这一小概率事件,给出了重叠碰撞拒绝检验;(2)基于大数定律,并通过多个样本构造样本方差,我们给出了无需已知理论方差的多样本 χ^2 分布检验.针对现有碰撞搜索方法在重叠统计方式下的资源占用高、搜索效率低等问题,我们提出了一种资源受限下基于CUDA的快速搜索方法,即通过前缀分类、后缀排序等优化技术,能够降低内存等硬件资源 $1/2^b$ 倍(其中 b 为前缀比特长度),同时也提高算法并行性,再利用CUDA并行技术,实现了重叠碰撞搜索在通用计算机上的实际可行性.我们利用提出的优化算法进行具体实验,在我们的实验平台上4200s左右可完成10GB数据的64比特以上的重叠碰撞搜索与碰撞信息存储.同时通过简单的数据构造实验展示了我们所提出的新方法的优点,即能够发掘现有NIST检验方法无法发现的长模板非随机因素.最后我们利用1000GB随机数据给出了1GB随机数据的长模板统计特征,也验证了本文所提出的检验方法的有效性.本文提出的重叠碰撞检验方法结合快速搜索工具,能够满足实际大数据量随机数测评需求,更好地保障密码系统安全性.

关键词 随机性统计检验;大样本;重叠碰撞;并行搜索;分类排序

中图法分类号 TP309 **DOI号** 10.11897/SP.J.1016.2023.01636

Statistical Tests on Overlapping Collisions with Large Samples

CHEN Dong-Yu^{1),2)} FAN Li-Min¹⁾ CHEN Hua¹⁾ WANG Jian^{1),2)} FU Yi-Fang^{1),2)}

¹⁾(Trusted Computing and Information Assurance Laboratory, Institute of Software, Chinese Academy of Sciences, Beijing 100190)

²⁾(University of Chinese Academy of Sciences, Beijing 100049)

Abstract The quality of the random number generated by the random number generator is directly related to the security of the cryptographic system, and the statistical hypothesis test for randomness is the commonly used method for the quality evaluation of random number, which plays an important role in the practice of cryptography. In recent years, the development of information technology has greatly improved the ability of big data processing. Accordingly, the statistical test for randomness is also facing the urgent needs of larger sample, stronger test ability and faster evaluation speed. Considering the applicability of existing test methods in large (GB-level) sample evaluation, this paper makes an in-depth study on tests for long template overlapping collision in large samples. First, we construct the statistic on the number of collision pairs. Based

收稿日期:2022-09-26;在线发布日期:2023-04-07. 本课题得到国家重点研究计划项目基金(2020YFA0309704)资助. 陈东昱,博士研究生,主要研究方向为随机数发生器设计与分析. E-mail: dongyu2019@iscas.ac.cn. 范丽敏(通信作者),博士,高级工程师,主要研究方向为侧信道分析与防护、密码检验. E-mail: fanlimin@iscas.ac.cn. 陈 华,博士,正高级工程师,主要研究方向为侧信道分析与防护、密码检验. 王 舰,博士研究生,主要研究方向为侧信道分析与防护. 付一方,硕士研究生,主要研究方向为随机数发生器设计与分析.

on this statistic, a new collision statistical model suitable for all counting methods is proposed through theoretical derivation. Then we derive the expected value of the overlapping collision statistic with the parameters of m -bit template of n -bit sequence. Based on this conclusion, we propose two statistical tests on the overlapping collisions with some theoretical analysis: (1) Based on the fact that the expected number of overlapping collisions can be used as the upper bound of the probability of collision occurrence, we use the nontrivial collision expectation that is small enough to reject the small probability event of a long collision occurrence, and introduce the overlapping collision rejection test; (2) Based on the law of large samples and constructing the sample variance from multiple samples, we provide the approximate t -distribution test with multiple samples that does not require a known theoretical variance. Considering the problems of high resource consumption and low search efficiency of existing collision search methods under the overlapping situation, we propose a fast search method based on CUDA with resource constraints, which can reduce hardware resources such as memory by $1/2^b$ times through optimization skills such as prefix classification and suffix sorting. These skills also improve the parallelism of the algorithm, and then we use the CUDA parallel technology to realize the practical feasibility of overlapping collision search on general computers. We take advantage of the proposed optimization algorithm to carry out specific experiments, which can complete the overlapping collision search and collision information storage of more than 64-bit templates of 10 GB data in about 4200 seconds. At the same time, through simple data construction experiments, the advantages of our proposed new test are demonstrated, that is, it can discover the non-random factors with long templates that cannot be detected by the NIST statistical tests for randomness. Finally, we use 1000 GB of random data to show the statistical distribution of long overlapping collision templates for 1 GB of random data, and also verify the effectiveness of the theoretical models and testing methods proposed in this paper. The overlapping collision tests proposed in this paper, combined with the fast search tool, can meet the actual evaluation needs of large samples, and better ensure the security of cryptographic systems.

Keywords statistical test for randomness; large sample; overlapping collision; parallel search; classify and sort

1 引言

密码技术是网络空间安全的基础支撑技术,它保障了信息的机密性、完整性、不可抵赖性等安全属性.而随机数则是密码技术中不可或缺的基础组件之一,如会话密钥、随机 Nonce 等.随机数发生器(Random Number Generator, RNG)是产生随机数的主要密码部件,一般可分为两类:伪随机数发生器(Pseudorandom Number Generator, PRNG)和真随机数发生器(True Random Number Generator, TRNG).伪随机数发生器一般由某一数学算法产生,在已知输入的情况下输出是确定的,因此也被称为确定性随机数发生器.通常 TRNG 都是基于一些不确定的随机物理现象,可以直接用作随机比特源

或者为 PRNG 产生随机种子,因此也被称为非确定性随机数发生器. RNG 输出的随机数的质量直接关系到密码系统的安全性,近年来由于随机数的错误使用导致密码系统失效的案例屡见不鲜,质量较差的随机数会造成严重的信息安全事故,例如 OpenSSL 伪随机数发生器有限随机数导致用户隐私泄露事件^[1]和带后门的双椭圆曲线确定性伪随机数发生器 Dual_EC_DRBG 导致互联网私密信息泄露事件^[2]等等.

在对随机数发生器的安全性评估中,统计假设检验是被广泛采用的评估手段之一.被广泛应用于随机数发生器测试^[3-7],以检验其产生的随机数的均匀随机性.统计检验假设检验是一种经典统计方法用于评估数据集是否符合给定假设(一般称之为原假设,记为 $\mathcal{H}_0$).用于 RNG 的随机数检验的原假

设是待检的 RNG 是理想的,即生成的随机数独立同分布,且都服从所期望的原假设分布.随机性统计检验旨在从各个数学角度来判断目标输出是否符合随机分布.多个国家和组织都给出了自己的随机数检验标准规范,以规范信息产品设计与质量,保障信息安全,例如美国的 NIST 标准^[4]、德国 BSI 的 AIS 标准^[6]、我国的随机数检验规范^[7]等等,其中多数采用了相似的随机性检验方法,并且近年来一直在更新完善.另外随机性检验也常用来评估一些密码算法的随机性,例如杂凑函数、分组密码等等^[8],以统计方式验证其与随机函数、置换的不可区分性.

由美国 NIST 提出的 SP800-22 统计检验标准^[4]在国际上应用与研究最为广泛,其最先发表于 2001 年,随后在 2010 年进行了版本修改. NIST SP800-22 共包含 15 个检验项目,包括频数检验、块内频数检验、游程检验、最长 1 游程检验、二元矩阵秩检验、离散傅里叶检验、非重叠模板匹配检验、重叠模板匹配检验、Maurer 通用统计检验、线性复杂度检验、重叠子序列检验、近似熵检验、累加和检验、随机游动检验、随机游动向量检验等,其对所有检验项目采用了相同的二级检验策略,即对 N 个不同的序列进行检验得到 N 个 p 值,并观察是否服从既定分布,包括通过率检验(检验 N 次实验通过该项目的比率)和均匀性检验(检验 N 个 p 值是否服从均匀分布).我国的随机性检验规范共包含 17 个检验项目,同样采用了二级检验方法.近年来,存储技术和计算效率的发展使得信息数据量急剧提升,这极大促进了信息产业的快速发展和产品迭代,从 3G 到 4G 再到 5G,人们生活已经逐步适应越来越多的信息和越来越快的处理速度,同时计算方法也得到了跨越发展.对应计算资源的发展和计算方法的进步,密码技术也迎来了新的更广泛的应用需求.从而对 RNG 的随机数统计检验也提出了更大数据量、更强检验能力和更快检验速度的需求.这一需求反映到统计检验规范,则是在保证检验方法可靠性的前提下待检数据可以达到大规模数据量,检验参数可达到更大规模.面对这一需求,我们一方面要重新审视规范标准中的统计检验方法,即在大样本大参数下检验方法的可靠性问题,另一方面要研究提出新型大样本随机数统计检验方法,从而加强检验力度.

在大样本随机性统计检验的可靠性方面,很多研究文章指出了现有随机数统计检验方法在一级检验(15 项基础检验)和二级检验两方面都存在误差问题,例如:Coron 和 Naccache^[9]指出不准确的近似

方差会导致 Maurer 通用统计检验相比理论上出现约 2.67 倍的拒绝偏差;Hamano 等人^[10]指出重叠模板匹配检验的统计量使用的是利用近似符合泊松分布计算得到的近似概率值,使得出现假阳性问题,并其通过构造全 1 模板的递归模型得到了准确概率值;Pareschi 等人^[11]对极大样本的二级检验可靠性问题进行了研究,指出并评估了一级检验 p 值误差导致的二级检验的明显假阳性问题;Chen 等人^[12]指出离散傅里叶检验方案中存在着检验准确性问题,当待检序列长度较长或条数较多时,以往的离散傅里叶检验方案中经常会出现 Type I 类错误,并通过构造卡方分布统计量,提出了一种新的针对长序列的离散傅里叶检验方案,新方案具有更好的稳定性.

本文主要关注新型大样本随机数统计检验方法方面的研究.在各随机数检验标准中,关于模板类的检验是非常重要的检验方法,例如:不相交检验^[6]、非重叠模板匹配检验^[13]、重叠模板匹配检验^[14]、近似熵检验^[15]、重叠子序列检验^[16]等等.以上检验都以模板为基础来构造统计值,并分为非重叠(如不相交检验、非重叠模板匹配检验)和重叠(如重叠模板匹配检验、重叠子序列检验)两种统计方式.以上模板类检验方法可以有效诊断出待检序列可能存在的不随机情况,但检验范围、检验规模及可靠性方面仍存在着一些问题,主要包括:重叠模板匹配检验和非重叠模板匹配检验一次检验仅关注一种限定模板,同时重叠模板匹配检验、近似熵检验在大样本下也存在可靠性问题^[10],重叠子序列检验所需的样本量是模板长度的指数级量,难以有效扩大检验参数规模,不相交检验较为适合大样本长模板的需求,但其为非重叠统计方式,因此本文主要关注序列重叠碰撞类检验方法的研究,期望能够得到具有较高检验规模的高效检验方法.

本文我们提出两种新型序列重叠碰撞检验方法,即重叠碰撞拒绝检验和多样本 t 分布检验.首先,给出了序列模板碰撞对数这一统计指标,即待检序列中所有模板对(任意两个位置的模板)形成碰撞的个数.然后通过构造碰撞对数统计量 ξ ,建立了统一的序列模板碰撞对数统计模型,问题根据具体统计方式可分为 3 类,即非重叠、重叠和循环统计.由此利用该模型容易求得 n 比特长 m 比特模板的碰撞对数的准确期望值.基于重叠统计方式下的碰撞对数统计量 ξ ,我们提出了两种具有新型序列重叠碰撞检验方法:(1)重叠碰撞拒绝检验,我

们利用理论期望值确定重叠方式下出现碰撞的概率上界,通过选择合适参数形成小概率事件的检验,从而给出一种高效的检验方法,例如针对 1 GB 数据搜索 80 比特重叠碰撞、10 GB 搜索 88 比特重叠碰撞、100 GB 搜索 96 比特重叠碰撞等等,若搜到碰撞则拒绝检验序列的随机性假设;(2)多样本 t 分布检验,在重叠方式下统计量方差未知的情况下,基于大样本正态假设,构造得到 t 分布统计量,从而形成检验策略,例如 1000 个 1 GB 样本的 56 比特重叠碰撞对数的分布符合性。这两种方法都实现了检验规模的极大提升,检验模板长度可达 56 比特以上,同时所需数据量也具实践性。具备了序列重叠碰撞检验理论,在实际测评面临的重大问题便是检验效率问题。现有的重叠碰撞搜索算法存在资源消耗大、搜索效率低的问题,即以 1 GB 搜索 64 比特重叠碰撞为例,其需要占用至少 64 GB 内存,当数据量上升到 100 GB 时,则通用计算机无法满足检验需求,另外缺少合适的并行算法来加速搜索效率,使得在时间上也缺少实际可行性。因此,我们针对这一问题,提出了一种资源受限下基于 CUDA^[17] 的快速搜索算法,基本思想是牺牲复杂度、提高并行性,将搜索模板分割为前缀和后缀数据,以前缀分类对后缀进行基数排序^[18],这样既降低了硬件资源占用,又提升了并行性。通过具体实验,我们用 24 GB 显存即可完成 10 GB 数据 64 比特以上模板的重叠碰撞搜索,而平凡搜索算法则至少需要 640 GB 的不现实内存,并且仅需 4200 s 左右便可完成搜索(1 GB 数据约 400 s 左右),能够满足重叠碰撞检验的实际高效测评需求。基于我们提出的重叠碰撞检验理论和快速搜索技术也可给出更多检验方法。

本文第 2 节介绍模板碰撞的基本概念和相关模板检验方法;第 3 节给出模板碰撞对数统计模型和重叠碰撞数的期望;第 4 节提出两种重叠碰撞检验方法;第 5 节提出基于 CUDA 技术的重叠碰撞快速搜索方法;第 6 节对全文进行总结。

2 背景知识

2.1 基本概念

首先,我们这里介绍模板碰撞的基础概念,包括几种碰撞概念、重叠与非重叠的统计方式。

定义 1. 模板. 我们将一段固定长度的比特串称之为一个模板,例如 m 比特模板 $P = (\epsilon_1, \dots, \epsilon_m)$, 其中 $\epsilon_i \in \{0, 1\}$ 。

关于模板碰撞的相关研究和统计检验已有多年的发展,然而关于碰撞的计数解释各有差异,所以这里需要先明确几种定义,以确保在检验时符合实际需求,具体定义如下。

定义 2. 碰撞. 在长序列模板统计中,若存在两个模板相同,则称这两个模板形成一个碰撞。例如模板 $P = (a_1, \dots, a_m)$ 和 $Q = (b_1, \dots, b_m)$, 若 $a_i = b_i$, $i = 1, \dots, m$, 则 P 和 Q 形成一个碰撞。

定义 3. 碰撞数. 在长序列模板统计中,若共有 n 个模板,其中不同模板个数为 n_1 , 则我们定义碰撞数为模板总数 n 与不同模板个数 n_1 的差,即碰撞数 $= n - n_1$ 。

定义 4. 碰撞对数. 在长序列模板统计中,若共有 n 个模板,我们定义碰撞对数为 n 个模板互相相等的个数,即碰撞对数 $= \sum_{i,j=1, i \neq j}^n I(c_i = c_j)$, 其中 $I(x) = \begin{cases} 0, & \text{若 } x=0 \\ 1, & \text{若 } x=1 \end{cases}$ 。

定义 5. 非平凡碰撞. 在长序列模板统计中,若存在两个模板形成碰撞,且这两个模板起始位置之前的 1 比特和终点位置之后的 1 比特都不相等,则称为非平凡碰撞。

定义 6. 非重叠统计. 在 N 长比特序列的 m 比特模板统计中,不允许模板相互重叠,则有 $n = \lfloor N/m \rfloor$ 个 m 比特模板,即按 m 比特分割该长序列进行统计。

定义 7. 重叠非循环统计. 在 N 长比特序列的 m 比特模板统计中,允许各模板重叠,则有 $n = N - m + 1$ 个 m 比特模板,即按 m 比特宽度依次后移 1 比特对该长序列进行统计。

定义 8. 重叠循环统计. 在 N 长比特序列的 m 比特模板统计中,允许各模板重叠且将前 $m-1$ 比特补充至序列尾部,则有 $n = N$ 个 m 比特模板,即序列尾部补充前 $m-1$ 比特数据,按 m 比特宽度依次后移 1 比特对该长序列进行统计。

我们通过一个实例来解释这几个定义。

例 1. 设有 12 比特序列 110110101101, 统计 4 比特模板的碰撞情况。则该序列有 3 个 4 比特非重叠模板: 1101, 1010, 1101; 9 个重叠非循环模板: 1101, 1011, 0110, 1101, 1010, 0101, 1011, 0110, 1101; 12 个重叠循环模板: 1101, 1011, 0110, 1101, 1010, 0101, 1011, 0110, 1101, 1011, 0111, 1110; 其中非重叠统计下碰撞数和碰撞对数都为 1; 重叠非循环统计下碰撞数为 4, 碰撞对数为 5, 其中有 1 个

4 比特非平凡碰撞;重叠循环统计下碰撞数为 5,碰撞对数为 7,其中没有 4 比特非平凡碰撞。

2.2 模板碰撞类检验

现有的模板检验统计理论成果丰富,对其概率问题的研究成果包括文献[14,19-21],采用了包括复合泊松方法、Goulden-Jackson 聚类方法、基于马氏链的方法等等,然而这些概率研究都是针对确定的几个模板.重叠子检验和近似熵检验是针对所有模板的频次情况检验.其基于每种模板出现概率构造卡方统计量.不相交检验是基于生日碰撞模型^[22]求得相对准确的概率值.下面我们简要介绍这些方法,说明其检验目标和适用范围。

非重叠模板匹配检验^[13].对固定长度的确定非周期模板出现频数进行检验,即统计模板 $B = (\epsilon_1^0, \dots, \epsilon_m^0)$ 在字符串中出现次数

$$W = \sum_{i=1}^{n-m+1} I(\epsilon_{i+k-1} = \epsilon_k^0, k=1, \dots, m).$$

通过将待检序列分割为 N 块,利用统计量 $\chi_{obs}^2 = \sum_{j=1}^N \frac{(W_j - \mu)^2}{\sigma^2}$ 进行卡方分布假设检验,其中 $\mu = \frac{n-m+1}{2^m}$, 方差为 $\sigma^2 = n \left[\frac{1}{2^m} - \frac{(2m-1)}{2^{2m}} \right]$. 该检验一个模板对应一个检验 p 值结果。

重叠模板匹配检验^[14].拒绝固定长度全 1 模板过多或过少的情况,即通过将待检序列分割为 N 块,每块长度为 M 比特长,利用符合泊松分布近似计算全 1 模板出现频次概率.文献[10]通过递归推导给出了更准确的概率结果.然后通过构造统计量 $\chi^2 = \sum_{i=0}^K \frac{(v_i - N\pi_i)^2}{N\pi_i}$ 进行卡方分布假设检验.该检验针对序列的全 1 模板,文献[4]指出该检验通过简单修改可适用于检验任何周期模板 B 的频次情况,注意在实际 NIST SP800-22 检验代码中存在两种概率选择,即复合泊松近似概率和准确的概率值,文献[12]中指出这两种概率选择在大样本下会存在完全不同的结果,近似概率的使用会直接导致假阳性问题,所以本文使用默认模板长度的准确概率值进行检验。

重叠子序列检验^[16].检验所有 2^m 种 m 比特长模板的频次分布情况,即令 $v_{i_1, i_2, \dots, i_m}$ 表示模式为 $(i_1, i_2, \dots, i_m)$ 的模板出现个数,统计值

$$\Psi_m^2 = \sum_{i_1, i_2, \dots, i_m} \frac{\left(v_{i_1, i_2, \dots, i_m} - \frac{n}{2^m} \right)^2}{\frac{n}{2^m}}.$$

则 $\nabla \Psi_m^2 = \Psi_m^2 - \Psi_{m-1}^2$ 服从自由度为 2^{m-1} 的 χ^2 分布。

该检验利用每种模板出现概率为 $1/2^m$ 的事实,构造了 Pearson 卡方统计量,同时为了保证可靠性,要求每种模板出现至少 5 次。

不相交检验^[6]. n 比特长序列,检验模板长度为 m 比特是否出现碰撞,基于生日碰撞原理可得其存在 m 比特碰撞的概率为 $1 - \exp\left(-\frac{(n/m)^2}{2^{m+1}}\right)$.

在 AIS31 中采用参数 $n = \frac{N}{b} = \frac{2^{16} \times 48}{48} = 2^{16}$, $m = 2^b = 2^{48}$, 则出现重复的概率为 $P = 1 - \exp(-2^{-17}) \approx 2^{-17}$. 可以看到在 $2^{16} \times 48$ 比特长随机串中,将其分割为以 48 比特为单位模板的随机串,共 2^{16} 个,出现 48 比特串模板重复的概率为 2^{-17} ,属于小概率事件,如果出现,则拒绝其为随机序列的假设。

从以上模板碰撞检验方法可以看到,对于大样本长模板的高效检验需求,重叠模板匹配检验在大样本下扩大模板长度参数存在可靠性问题,且其概率参数准确计算复杂(难度随样本量增大而急剧增加),重叠子序列检验随着模板长度增加所需统计计算量指数级增大(每种模板至少出现 5 次),非重叠模板匹配检验适用模板长度短且只能逐个模板检验,不相交检验为非重叠统计方式的碰撞检验.因此我们自然想要考虑力度更大的重叠碰撞检验,即以重叠方式检验模板的碰撞性质,以 n 长 4 比特为例,非重叠方式有 $n/4$ 个模板,重叠方式则有 $n-3$ 个 4 比特模板.然而目前缺少相应的检验理论与方法,基于此,本文考虑提出更适用于大样本的长模板重叠碰撞类检验方法,满足对大样本以重叠方式进行长模板碰撞类检验的实际测评需求。

3 碰撞统计模型

不相交检验可通过生日碰撞模型求出一个极小碰撞概率,来拒绝出现碰撞的情况,然而理论上计算重叠统计下的碰撞概率实属困难,对于不同的模板其递推情况不一致,依赖关系相当复杂,现在的研究都是集中在一个模板或多个模板的频次检验.这里,我们考虑用碰撞对数构造统计量,进而推导重叠碰撞数的期望模型,并给出一种简单计算方法.构建如下概率统计模型:

模型: 序列 $S \in \{0,1\}^n$, 即为 n 长 01 序列;模板 $P \in \{0,1\}^m$, 为 m 长 01 子序列,共有 2^m 种。

问题: 01 序列中的碰撞模板对的个数 ξ ?

问题分类: 首先,问题存在 3 种分类,即:

(1) 非重叠模板统计,即在对序列 S 中模板进

行统计时,考虑各模板不重叠情形,则共有 $\lfloor n/m \rfloor$ 个模板;

(2)重叠模板非循环统计,即在对序列 S 中模板进行统计时,考虑各模板可重叠但不补充情形,则共有 $n-m+1$ 个模板;

(3)重叠模板循环统计,即在对序列 S 中模板进行统计时,考虑各模板可重叠并补充前面比特情形,共有 n 个模板.

注意,该统计模型中关注的是碰撞对数,而不是我们经常遇到的碰撞数,碰撞对数 $\geq$ 碰撞数=所有模板数-不同模板数.举个简单例子:设 $A=B=C$,则有 3 个碰撞对,2 个碰撞.

定理 1. 对以上三类问题,都成立:设共有 k 个模板,则序列 S 的碰撞对数期望为 $\binom{k}{2}/2^m$.

证明. 共有 k 个模板,则有 $\binom{k}{2}$ 对模板需要比较以确定碰撞对数.令 I_q 是表示第 q 对模板(即模板 P_{q_i} 与 P_{q_j})是否相等的一个随机变量,即:

$$I_q = \begin{cases} 1, & \text{若 } P_{q_i} = P_{q_j}, \\ 0, & \text{否则} \end{cases}, \text{ 其中 } q=1, \dots, \binom{k}{2}$$

设 $P=(a_1, \dots, a_m)$ 和 $Q=(b_1, \dots, b_m)$ 为序列中任意两个模板,则:

$$\text{Prob}(P=Q) = \text{Prob}(a_1=b_1, \dots, a_m=b_m).$$

若 P 和 Q 不重叠,则显然:

$$\begin{aligned} \text{Prob}(P=Q) &= \text{Prob}(a_1=b_1) \cdot \dots \cdot \text{Prob}(a_m=b_m) \\ &= \left(\frac{1}{2}\right)^m. \end{aligned}$$

若 P 和 Q 重叠:当重叠 $m-1$ 比特时,两模板相等则相当于 $m+1$ 个连续 0 或 1 比特的情形, $\text{Prob}(P=Q) = 2/2^{m+1} = 1/2^m$;当重叠 $m-2$ 比特时,两模板相等则相当于 $m+2$ 比特串为“ $abababab\dots$ ”情形(其中 $a, b \in \{0, 1\}$),则 $\text{Prob}(P=Q) = 4/2^{m+2} = 1/2^m$;以此类推,当重叠 1 比特时,两模板相等则相当于 $2m-1$ 比特串为“ $a_1, \dots, a_m, b_2, \dots, b_m$ ”情形(其中 $a_i, b_i \in \{0, 1\}$),则 $\text{Prob}(P=Q) = 2^{m-1}/2^{2m-1} = 1/2^m$.

因此,在各统计方式下,任意两个 m 比特的模板相等的概率皆为 $1/2^m$. 则第 q 个模板对的期望和方差如下:

$$E(I_q) = \text{Prob}(P_{q_i} = P_{q_j}) = \frac{1}{2^m},$$

$$D(I_q) = E(I_q^2) - (E(I_q))^2 = \frac{1}{2^m} - \left(\frac{1}{2^m}\right)^2.$$

令 $\xi = \sum_{q=1}^{\binom{k}{2}} I_q$, 为所有 $\binom{k}{2}$ 个模板对中出现碰撞的个数,即序列的碰撞对数,则我们容易得到,序列 S 的

模板碰撞对数期望为

$$E\xi = E\left(\sum_{q=1}^{\binom{k}{2}} I_q\right) = \sum_{q=1}^{\binom{k}{2}} E(I_q) = \binom{k}{2} / 2^m. \text{ 证毕.}$$

注意到定理 1 并不针对某一种统计方式,因此该结论适用所有统计方式,考虑将各统计方式下的模板数代入,容易得到如下推论.

推论 1. 各类问题的碰撞对数期望为

情形(1): $k = \lfloor n/m \rfloor, E\xi = \binom{k}{2} / 2^m$;

情形(2): $k = n-m+1, E\xi = \frac{(n-m+1) \cdot (n-m)}{2^{m+1}}$;

情形(3): $k = n, E\xi = \frac{n(n-1)}{2^{m+1}}$.

容易看到情形(2)即为本文所关注的重叠碰撞情形,我们对其进行了小规模随机实验,验证其期望结论的正确性.如图 1 所示,随机生成 $n=100$ 比特序列,我们寻找 2 到 20 比特重叠碰撞对数,结果显示 1000 次随机实验均值与理论期望值 $\frac{(n-m+1) \cdot (n-m)}{2^{m+1}}$ 基本一致;如图 2 所示,随机生成 50 到 200 比特长度的序列,寻找 $m=10$ 比特的重叠碰撞对数,进行 1000 次随机实验,结果依然显示理论期望值与随机实验均值基本一致.

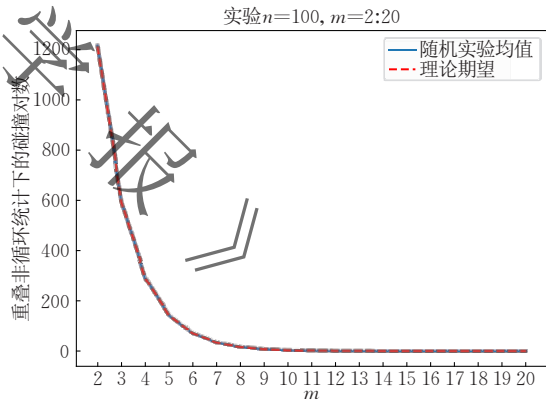


图 1 重叠非循环统计下碰撞对数期望实验 1

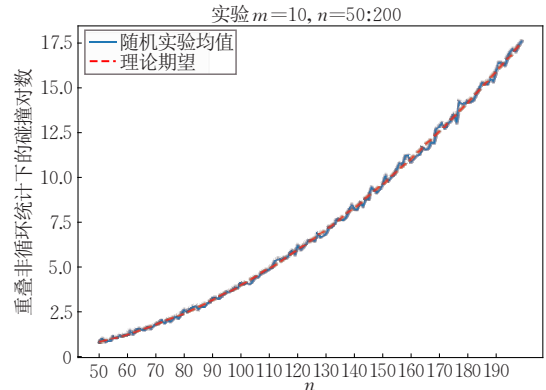


图 2 重叠非循环统计下碰撞对数期望实验 2

在测评实践中经常会考虑最长模板碰撞,而非平凡碰撞表示各位置的最长碰撞模板,利用同样的技术原理,我们也可以给出非平凡碰撞对数的相关结论(以重叠非循环统计方式为例)。

推论 2. n 比特长序列 S , 以重叠非循环方式统计, 有 $n-m+1$ 个 m 比特模板, 则非平凡碰撞对数的期望为 $\frac{(n-m-1)(n-m-2)}{2^{m+3}} + \frac{n-m}{2^m}$ 。

证明. 设共有 k 个模板, 则有 $\binom{k}{2}$ 对模板需要比较以确定非平凡碰撞对数. 令 I_q 是表示第 q 对模板 (即模板 P_{q_i} 与 P_{q_j}) 为如下随机变量:

$$I_q = \begin{cases} 1, & \text{若 } P_{q_i} \text{ 与 } P_{q_j} \text{ 为非平凡碰撞} \\ 0, & \text{否则} \end{cases}$$

其中 $q=1, \dots, \binom{k}{2}$. 则 $I_q=1$ 的概率为“模板 $P_{q_i}=P_{q_j}$ 且两模板在序列中的前后 1 比特不相等”的概率, 即有:

(1) 当 P_{q_i}, P_{q_j} 都不为序列第一个和最后一个模板时,

$$\text{Prob}(I_q=1) = \frac{1}{2^m} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{2^{m+2}}.$$

(2) 当 P_{q_i}, P_{q_j} 中一个为序列第一个或最后一个模板时, 则模板在序列中无前或后 1 比特数据, 即:

$$\text{Prob}(I_q=1) = \frac{1}{2^m} \cdot \frac{1}{2} = \frac{1}{2^{m+1}}.$$

(3) 当 P_{q_i}, P_{q_j} 分别为序列第一个和最后一个模板时, 则模板在序列中无前和后 1 比特数据, 即:

$$\text{Prob}(I_q=1) = \frac{1}{2^m}.$$

由此, 令 $\xi = \sum_{q=1}^{\binom{k}{2}} I_q$, 即为序列的碰撞对数, 则序列 S 的模板碰撞对数期望为

$$\begin{aligned} E\xi &= E\left(\sum_{q=1}^{\binom{k}{2}} I_q\right) = \sum_{q=1}^{\binom{k}{2}} E(I_q) \\ &= \binom{k-2}{2} \cdot \frac{1}{2^{m+2}} + 2 \cdot (k-2) \cdot \frac{1}{2^{m+1}} + \frac{1}{2^m} \\ &= \frac{(k-2)(k-3)}{2^{m+3}} + \frac{k-1}{2^m}. \end{aligned}$$

重叠非循环统计下, $k=n-m+1$, 则非平凡碰撞对数的期望为

$$\frac{(n-m-1)(n-m-2)}{2^{m+3}} + \frac{n-m}{2^m}. \quad \text{证毕.}$$

显然, 通过统计非平凡重叠碰撞对数, 可以得到重叠碰撞对数, 实践中这种统计方式能够更好地展示碰撞结果, 同时更易于重叠碰撞搜索算法的优化

设计. 同时注意到, 重叠方式下各模板的依赖关系较为复杂, 依赖情形依模板而不同, 然而这种依赖关系并不影响期望的计算, 所以我们给出了重叠碰撞的准确理论期望值, 然而其理论方差则未能得到显式表达, 但我们依然可以通过实验设计得到实用的随机性检验方法。

4 重叠碰撞检验

本节中, 我们设计了两种序列重叠碰撞类检验方法, 包括重叠碰撞拒绝检验和多样本 t 分布检验, 注意我们主要考虑重叠非循环情形, 下面均简称为重叠统计方式。

4.1 重叠碰撞拒绝检验

注意到第 3 节中的情形(1)与前面介绍的不相交检验模型非常相似, 不相交检验关注是否出现碰撞, 情形(1)则关注出现的碰撞对数, 容易看到碰撞对数 ≥ 1 等价于出现碰撞. 将不相交检验的参数代入情形(1), $k=2^{16}$, $m=48$, 则碰撞对数期望为 $E\xi = \frac{2^{16} \cdot (2^{16}-1)}{2 \cdot 2^{48}} = 2^{-17} - 2^{-33} \approx 2^{-17}$. 可以看到, 在出现碰撞的概率极小时, 碰撞对数期望值与碰撞概率很接近。

基于以上观测, 我们继续讨论重叠统计下碰撞概率, 尝试去估计该概率. 我们这里寻求 m 比特长度重叠碰撞出现的概率极小, 也即序列的最长重叠碰撞的长度超过 m 比特的概率极小, 注意到最长重叠碰撞一定是一个非平凡重叠碰撞. 首先, 依据推论 2, 我们可以很轻松得到出现非平凡碰撞对数的期望:

$$E\xi = \frac{(n-m-1)(n-m-2)}{2^{m+3}} + \frac{n-m}{2^m}. \quad \text{那么成立:}$$

$$E\xi = 0 \cdot p(0) + 1 \cdot p(1) + \dots + i \cdot p(i) + \dots,$$

其中 $p(i)$ 表示出现 i 个非平凡重叠碰撞对的概率. 我们要求的是出现 m 比特非平凡重叠碰撞的概率:

$$p = 1 - p(0) = p(1) + p(2) + p(3) + \dots,$$

则显然成立:

$$p < E\xi = \frac{(n-m-1)(n-m-2)}{2^{m+3}} + \frac{n-m}{2^m}.$$

即我们这里计算的期望值给出了存在 m 比特非平凡重叠碰撞的概率上界. 我们可以选择理论期望值足够小的参数值 n, m , 从而可知 n 比特长序列出现 m 比特非平凡重叠碰撞属于小概率事件, 由此来进行随机性检验. 类似不相交检验, 我们选取 $E\xi$ 足够小的检验参数 (碰撞模板长度), 得到一种重叠碰撞拒绝检验. 基于这一点, 我们可以给出重叠碰撞

拒绝检验策略,即利用该期望公式,选择“出现某长度的重叠模板碰撞”的极小概率事件,若出现该事件则拒绝该序列为随机序列的假设.具体如下.

重叠碰撞检验模型. n 比特长随机序列,其最长重叠碰撞的长度超过 m 比特的概率上界(碰撞对数期望)为

$$\frac{(n-m-1)(n-m-2)}{2^{m+3}} + \frac{n-m}{2^m}.$$

由此可得表 1 所示各参数规模的结果,其中列为待检序列长度,行为固定长度模板的长度,表中值为碰撞对个数的(以 2 为底的)对数值:

$$\log\left(\frac{(n-m-1)(n-m-2)}{2^{m+3}} + \frac{n-m}{2^m}\right).$$

表 1 重叠碰撞模型参数示例

	64 比特	72 比特	80 比特	88 比特	96 比特
1 GB	-1.00000	-9.00000	-17.00000	-25.0000	-33.000
2 GB	0.99999	-7.00000	-15.00000	-23.0000	-31.000
10 GB	5.64386	-2.35614	-10.35610	-18.3560	-26.356
100 GB	12.28877	4.28771	-3.71228	-11.7122	-19.712

由表 1 可知,1 GB 数据 80 比特重叠碰撞对数期望为 $2^{-17.000}$,则 1 GB 数据出现 80 比特重叠碰撞的概率小于 $2^{-17.000}$,由此我们可以对 1 GB 序列进行 80 比特重叠碰撞的拒绝检验,即若 1 GB 数据以重叠方式统计,搜索得到 80 比特的碰撞,则拒绝该序列为随机序列的假设,即判断该序列不是随机序列.由此,表 1 可以作为我们的拒绝检验策略,例如 1 GB、2 GB 搜索 80 比特重叠碰撞,10 GB 搜索 88 比特重叠碰撞,100 GB 搜索 96 比特重叠碰撞,若搜得碰撞则拒绝检验序列的随机性假设.以上,我们为计数方便确立的搜索碰撞长度为 8 的倍数,当然可以通过重叠碰撞模型得到更为紧致的搜索策略,例如需满足 $\log\left(\frac{(n-m-1)(n-m-2)}{2^{m+3}} + \frac{n-m}{2^m}\right) < -10$,即理论上 n 比特长序列最长重叠碰撞的长度超过 m 比特的概率小于 0.000977,例如 1 GB 查 73 比特碰撞,10 GB 查 80 比特碰撞,100 GB 查 87 比特碰撞等等.

4.2 多样本 t 分布检验

当样本量较大时,我们可以考虑用近似正态分布来进行重叠碰撞对数统计量的检验,然而其理论方差不易求得.为了避免使用不准确的方差参数而影响检验可靠性,我们考虑直接用理论期望值构造统计量进行假设检验,即通过多样本构造样本方差形成学生 t 分布统计量^[23].

多样本 t 分布检验模型. 设 n 比特长序列的 m 比特重叠碰撞对数的理论期望为 $a = ((n-m+1) \cdot$

$(n-m))/2^{m+1}$,理论方差为 σ^2 ,取 N 个 n 比特长随机序列(相当于进行 N 次统计实验),令 $\bar{\xi} = \sum_{i=1}^N X_i / N$ 为样本的重叠碰撞对数均值,其中 X_i 为第 i 个 n 比特样本的重叠碰撞对数,则对于较大的 N 和 n ,成立如下近似:

$$\frac{\bar{\xi}-a}{\sigma/\sqrt{N}} \sim \mathcal{N}(0,1).$$

定义 $S^2 = \frac{1}{N} \sum_{i=1}^N (X_i - \bar{\xi})^2$ 为 n 比特长随机序列

m 比特重叠碰撞对数的样本方差,则 $\frac{NS^2}{\sigma^2}$ 服从自由度为 $N-1$ 的卡方分布,即:

$$\frac{NS^2}{\sigma^2} \sim \chi^2(N-1).$$

由此可得, $\frac{\bar{\xi}-a}{S/\sqrt{N-1}}$ 服从自由度为 $N-1$ 的 t 分布,即:

$$\frac{(\bar{\xi}-a)/(\sigma/\sqrt{N})}{\sqrt{NS^2/\sigma^2/(N-1)}} = \frac{\bar{\xi}-a}{S/\sqrt{N-1}} \sim t(N-1).$$

最终我们得到了一种重叠碰撞检验方法,其显著性水平为 α 的检验策略为:

若 N 个 n 比特长样本序列的 m 比特长模板的重叠碰撞对数均值 $\bar{\xi} = \sum_{i=1}^N X_i / N$ 在如下区间内 $\left[a - \frac{S \cdot t_{1-\frac{\alpha}{2}}(N-1)}{\sqrt{N-1}}, a + \frac{S \cdot t_{1-\frac{\alpha}{2}}(N-1)}{\sqrt{N-1}} \right]$,

则接受检验序列为随机序列的假设;否则,拒绝.

下面,我们进行随机实验用以验证 $\frac{\bar{\xi}-a}{S/\sqrt{N-1}}$ 近似服从 $t(N-1)$ 分布,具体内容为针对 100 比特长 5 比特模板的重叠碰撞对数,进行了 1000 次均值实验,每次实验随机采取了 1000 个 100 比特样本,结果如图 3 所示.我们以 20 个区间分组对 1000 个实验结果的标准化值 $\frac{\bar{\xi}-a}{S/\sqrt{N-1}}$ 进行直方图统计,其中每个实验结果 $\bar{\xi}$ 、 S 由 1000 个样本计算而得.由图 3 可见,近似 $t(N-1)$ 分布的实验数据拟合效果较好,符合理论结论.

注意到,如果参数 m 过大,其期望值较小,不存在 m 比特重叠碰撞的概率较大,则实际重叠碰撞对数显然不符合近似正态分布的对称性(如图 6).所以需要谨慎选择统计模板长度,例如我们这里给出一种实际检验策略:1000 GB 数据分割为 1000 份样本,每个 1 GB 样本(即 $n=2^{33}$)搜索 $m=56$ 比特重叠

碰撞对数,计算检验统计量 $\bar{\xi}$,若不在检验策略区间 $\left[a - \frac{S \cdot t_{1-\frac{\alpha}{2}}(N-1)}{\sqrt{N-1}}, a + \frac{S \cdot t_{1-\frac{\alpha}{2}}(N-1)}{\sqrt{N-1}} \right]$ (其中 $a = \frac{(n-m+1) \cdot (n-m)}{2^{m+1}}$, $N=1000$) 内则拒绝为随机序列的假设 (其中 α 取 0.01).

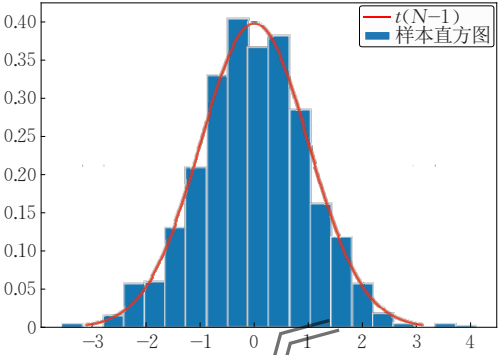


图 3 近似多样本 t 分布的随机实验

4.3 讨论

本节给出了两种 n 比特长序列的 m 比特模板的碰撞对数检验方法:重叠碰撞拒绝检验和多样本 t 分布检验.重叠碰撞拒绝检验采用了一种简单的检验策略,即出现某长度的重叠碰撞则拒绝,是对可能出现的最大重叠碰撞长度的检测,可适应各种样本量的检验,可检验模板长度较长,非常适合高效测评需求.多样本 t 分布检验在期望值可知而方差未知的情况下构造了 t 分布统计量,其更关注固定长度的重叠碰撞模板频数的随机符合性,相对于拒绝检验是一种更为深入的随机性检验,对同一检验模板长度,其所需样本量相对上述拒绝检验要大很多,但与重叠模板匹配检验、重叠子序列检验相比,则大大降低了所需样本量 (例如重叠子序列检验要求所有模板出现至少 5 次,这对于像 56 比特这种长模板来说统计难度过大),能够满足大样本统计测评的理论保障性和实际可行性.

因此,我们希望大样本重叠碰撞类统计检验能够进一步降低样本量需求,如果可知理论方差,则可以取得较大进步.同时我们认为即使只得到理论方差较紧的上界估计,依然可以获得较好的检验策略.例如,根据车贝晓夫不等式:

$$P\{|\xi - E\xi| \geq \epsilon\} \leq \frac{D\xi}{\epsilon^2},$$

我们可以得到一种普适的检验策略,即不论 ξ 的准确分布是什么, ξ 落在 $(E\xi - \delta\sigma, E\xi + \delta\sigma)$ 中的概率不小于 $1 - 1/\delta^2$. 因此对 $D\xi$ 的上界估计或误差有界的估计都可以得到一种高效的检验方法.例如可采取策略有:对难以分析的模板重叠情形采取上界

估计,得到近似理论方差估计;利用优质随机数通过大数据实验给出样本方差的最优无偏估计 $S_*^2 = \frac{1}{N} \sum_{i=1}^N (X_i - a)^2$, 其中 a 为重叠碰撞期望,同时基于卡方分布给出方差的 $1-\alpha$ 置信区间估计:

$$\left\{ \sigma^2: \frac{(n-1) \cdot S_*^2}{\chi_{n-1, 1-\frac{\alpha}{2}}^2} \leq \sigma^2 \leq \frac{(n-1) \cdot S_*^2}{\chi_{n-1, \frac{\alpha}{2}}^2} \right\}$$

等等,我们在之后的工作中将结合大数据量的实验对其进行深入研究.

5 重叠碰撞快速搜索方法

本节主要关注 n 比特 (1GB~100GB) 长序列的 m 比特 (64 比特以上) 长度的重叠碰撞搜索,我们给出了优化实现方法,能够满足高效检验需求.

5.1 平凡实现

首先,最平凡的实现方法是:一般实现.直接将数据读入内存,按 m 比特重叠模板进行比较.复杂度为 $O(n^2)$. 以 1GB 查 72 比特重复为例,该实现方式至少需要 72GB 内存;当数据量达到 10GB 时,则至少需要 720GB 内存;100GB 时,至少需要 7200GB 内存.显然这种实现方式需要超大内存,不适合通用计算机,且其本身逐对对比的复杂度也较高.为了加快实现速度,较为高效的方法是:快排实现.直接将数据读入内存,按 m 比特重叠模板快排再比较.复杂度 $O(n \log n)$. 相比一般实现的方式,该实现方式大大降低了复杂度,但所需内存依然不变,这种实现方式更适合高性能计算机,然则其对内存的极高需求也极大降低了高性能计算机的数据量适用范围.通过利用计算机的并行能力 (CPU 和 GPU),显然可以极大加速搜索,然而这并没有改变计算资源的限制,从而极大限制了可检验数据量.因此,在实际工作中,为了兼顾计算资源和效率,急需高效的重叠碰撞搜索方法,而这无疑加大了搜索算法的设计难度.

5.2 资源受限下基于 CUDA 的快速搜索方法

为了解决上述存在的计算资源受限和效率低的问题,即解决难以在计算资源受限下高效搜索重叠碰撞的问题,本节提出一种资源受限下基于 CUDA 的序列碰撞快速搜索方法,能够在有限计算资源下进行重叠碰撞的高效搜索,极大提高检验数据量上限,从而更好地服务于 RNG 随机数质量检验工作.

本节主要优化资源占用和搜索效率两个方面: (1) 将待检数据进行分类匹配,再进行碰撞搜索,通过缩小搜索空间来减少资源占用; (2) 利用基于计算机 GPU 的 CUDA 并行技术,提高搜索效率.本节

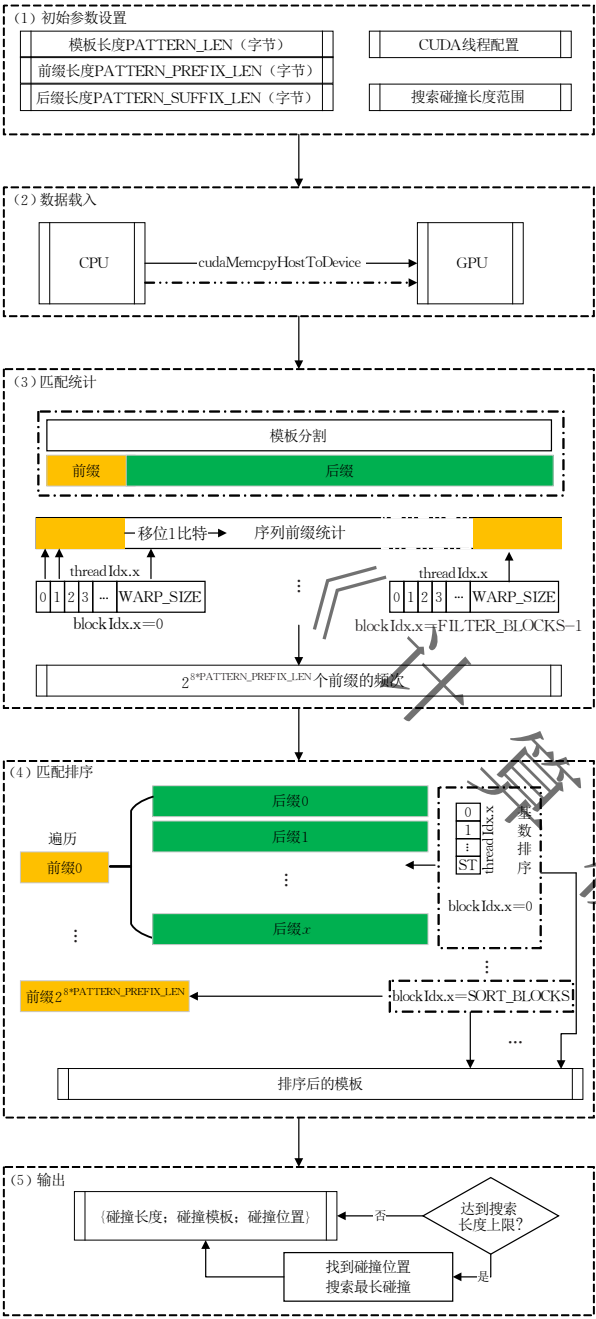


图 4 资源受限下基于 CUDA 的重叠碰撞搜索算法

所提出的资源受限下基于 CUDA 的序列碰撞快速搜索方法参见图 4,具体流程描述如下。

算法 1. 资源受限下基于 CUDA 的重叠碰撞搜索。

输入: n 比特长序列 S , 重叠碰撞模板长度范围 $[m_1, m_2]$ (比特)

输出: 所有以上长度范围内的重叠碰撞码字、长度和位置

1. 初始参数设置. 依据计算资源和 $[m_1, m_2]$, 设定检查模板长度 PATTERN_LEN (字节, $\leq \lceil m_2/8 \rceil$)、前缀长度 PATTERN_PREFIX_LEN (字节)、后缀长度 PATTERN_SUFFIX_LEN (字节). 注意 PATTERN_LEN = PATTERN_

PATTERN_LEN + PATTERN_SUFFIX_LEN. 设定 CUDA 并行线程 (线程块、线程束), 设定输出碰撞模板长度范围 (注: 下限 m_1 过小不具意义, PATTERN_LEN 过大影响算法效率和资源过度占用, 需依据理论期望结果考虑)。

2. 数据载入. 将待检数据加载到 GPU 设备. 首先通过 CPU 将待检数据读入, 然后复制 (cudaMemcpyHostToDevice) 到 GPU 设备. 对于超大数据 (超出 GPU 显存负载), 可采用多次数据传输, 即 CPU 与 GPU 进行多次传输交互, 完成下面的流程。

3. 匹配统计. 对数据进行 CUDA 并行匹配, 得到所有 PATTERN_PREFIX_LEN 字节前缀的频次。

4. 匹配排序. 利用 CUDA 并行, 遍历所有前缀, 通过匹配前缀得到相应的所有后缀数据, 并对其进行排序, 得到所有满足长度要求的碰撞及其所出现位置和长度。

4.1. 遍历前缀. 依据前缀频次、设备显存大小、并行线程数, 统筹一次处理 (≥ 1) 多个前缀对应的后缀数据的排序。

4.2. 后缀排序. 采用更适合 CUDA 并行的排序算法——基数排序, 即并行首字节 (或多字节, 依并行线程分配而定), 后逐字节块通过偏移量统计进行排序。

5. 输出. 将结果传输到 CPU 端进行结果处理并输出 {碰撞长度; 碰撞模板; 碰撞位置}, 即若碰撞比特长度 $\geq m_1$ 则进行标准输出; 若 $8 * \text{PATTERN_LEN} < m_2$, 则直接在 CPU 端对碰撞长度 = PATTERN_LEN 的碰撞位置进行更长碰撞 (直至 m_2 比特长度) 的判定搜索。

与前面平凡实现方法相比, 我们的搜索方法显著降低了计算资源消耗, 并极大提升了搜索效率. 针对重叠碰撞搜索超大资源占用问题, 提出了前缀分类匹配搜索的方法, 即将模板分为前缀和后缀模板, 匹配前缀数据对后缀数据进行碰撞搜索. 假设设定前缀为 b 比特, 则可实现搜索阶段资源占用至少降低 $1/2^b$ 倍. 显然该方法牺牲了复杂度, 但却极大提高了并行性, 可有效利用 CUDA 并行技术, 对后缀数据进行排序搜索碰撞, 从而实现高效的碰撞搜索. 同时对于超长碰撞并不直接进行搜索, 而是依据计算资源和理论期望计算设定搜索长度上限, 对于达到上限长度的碰撞再进行进一步的搜索, 这样既可以极大降低超长模板重叠统计的资源占用 (依据理论期望计算, 会选则期望数较小的碰撞长度作为上限, 在此基础上继续搜索更长碰撞, 增加的搜索时间极大概率意义下可忽略), 又可以避免牺牲计算性能 (在排序阶段超长模板排序会造成排序效率的降低, 算法层面上增加了迭代深度)。

6 实验分析

利用我们提出的重叠碰撞搜索的优化算法, 我

们依据所用计算资源通过具体 CUDA 程序实现,进行了系列具体实验.需要注意的是,碰撞统计方式在搜索方法上并没有什么区别.

实验 1. 10 GB 数据搜索效果测试.

我们以 10 GB 待检验数据为例,搜索 64~128 比特碰撞模板、位置、长度等.注意到通过表 1 可知,对于 1 GB~100 GB 数据,64 比特以上碰撞数理论期望值皆较小,属于合理实验范围.我们具体实验所用计算机配置为 Intel Xeon 金牌 5218,16 核,32 线

程,2.3 主频,32 GB 内存,显卡 RTX 3090 24 GB,电源 1400 W.针对 10 GB 数据,这里我们搜索重叠非平凡碰撞,依计算资源情况设置搜索算法的关键参数,具体实现参数可见附录 A.如表 2 所示,我们进行了多次实验,皆约在 4200 s 左右可以完成对 10 GB 数据的所有 64~112 比特重叠碰撞模板及其位置和长度的输出.具体碰撞模板码字和位置搜索结果,以表 2 中数据 1 的 68~70 比特重叠碰撞的结果为例,见表 3.

表 2 10 GB 重叠非平凡碰撞搜索示例(搜索结果 为各碰撞长度下的重叠非平凡碰撞数)

10 GB 测试集	搜索时间/s	64 比特	65 比特	66 比特	67 比特	68 比特	69 比特	70 比特	71 比特	72 比特
数据 1	4133	106	56	31	12	6	1	1	0	0
数据 2	4389	82	46	20	10	3	0	0	0	0
数据 3	4342	103	48	29	15	6	3	1	1	1
数据 4	4066	98	54	31	19	8	2	0	0	0

表 3 10 GB 数据重叠碰撞的详细搜索结果示例

长度	碰撞位置	碰撞码字
68	{0000000937079040, 0000073486689012}	{0x02385566ce84af2b8eae6464d850, 0x02385566ce84af2b84f616930edb}
68	{0000021884714354, 0000022004136187}	{0x3486468dcc8bb394c49a83179581, 0x3486468dcc8bb394cbbe116622ad}
68	{0000053730695714, 0000067916430898}	{0x9ab4fa3c1dc1c0e7092b508790e5, 0x9ab4fa3c1dc1c0e707ea84390513}
68	{0000000492096837, 0000031159921173}	{0xa566271c06ddcdca12d3cdf77d62, 0xa566271c06ddcdca1bf631b8d3b4}
68	{0000007918158459, 0000075198177275}	{0xb24bfc0936539c35e4fb555abfd5, 0xb24bfc0936539c35ed5254e18db8}
68	{0000061383859725, 0000064240235163}	{0xc879a9f288fae774102272a36583, 0xc879a9f288fae7741f0a4b23b58a}
69	{0000007918158458, 0000075198177274}	{0xd925fe049b29ce1af27daad5fea, 0xd925fe049b29ce1af6a92a70c6dc}
70	{0000007918158457, 0000075198177273}	{0xec92ff024d94e70d793ed556aff5, 0xec92ff024d94e70d7b549538636e}

由此实验可以看到,我们的搜索方法能够在较为通用的计算平台上,完成计算资源受限下重叠碰撞的快速搜索,可满足实际大数据高效检验需求.

实验 2. 重叠碰撞拒绝检验实验.

我们通过此实验验证,第 4 节提出的检验方法相比现有的随机数检验方法是否能够发现更深层次的非随机因素.我们采用 AES-CBC 算法生成 1 GB

随机数据,同时随机选取两个位置构造 80 比特重叠碰撞,最终该数据可以通过 NIST SP 800-22 随机数检验(通过率大于 0.981,均匀度大于 0.0001),但明显无法通过提出的重叠碰撞拒绝检验.显然这样的实验样本很容易构造,同时也很容易成功.我们构造实验示例得到如表 4 中所示的结果(注:表 4 中我们使用全 0xff 的密钥,全 0x00 的明文输入,在第 368 字

表 4 具有 80 比特重叠碰撞的 1 GB 数据随机性检验结果

检验项目	$n=1\,000\,000, N=1000$		$n=2^{23}, N=2^{10}$	
	通过率检验	均匀性检验	通过率检验	均匀性检验
频数	0.984 000	0.286 836	0.988 281	0.972 236
块内频数	0.990 000	0.102 526	0.986 328	0.512 565
游程	0.983 000	0.410 055	0.990 234	0.333 308
最长 1 游程	0.989 000	0.191 687	0.985 352	0.608 563
矩阵秩	0.987 000	0.605 916	0.992 188	0.600 424
离散傅里叶	0.986 000	0.055 010	0.992 188	0.283 859
非重叠模板匹配	0.990 000	0.410 055	0.994 141	0.978 981
重叠模板匹配	0.990 000	0.837 781	0.989 258	0.770 860
通用统计	0.982 000	0.440 975	0.987 305	0.628 954
线性复杂度	0.992 000	0.80 1865	0.990 234	0.967 323
重叠子序列	0.989 000	0.731 886	0.992 188	0.118 880
近似熵	0.989 000	0.208 837	0.990 234	0.653 447
累加和	0.984 000	0.548 314	0.986 328	0.253 551
随机游动	0.990 369	0.875 539	0.986 143	0.548 267
随机游动变量	0.988 764	0.116 153	0.990 762	0.129 968
80 比特重复	9b2e907fff750c6f2ccc		9b2e907fff750c6f2ccc	

节和 3680 字节处构造重复),由此结果可知,我们提出大样本长模板的重叠碰撞检验策略能够找到现有 NIST 检验方法无法发现的非随机因素,因为现有 NIST 检验方法所针对的模板对象较短,长模板碰撞不会造成现有检验方法统计量的明显差异,因此该策略完全可以作为现有检验标准的强力补充,同时其本质上是在样本量和检验规模上的提升,提供了更深层次的检验.注意,正如我们在前言中所说,重叠模板匹配检验默认采用近似概率计算,从而样本量变大会得到不可靠的结果,所以我们这里(1 GB)使用了准确计算的概率值.

实验 3. 1GB 样本的均值和方差实验.

我们利用 AES-CBC 算法产生了 1000 GB 随机数,用来观察理论均值

$$a = \frac{(n-m+1) \cdot (n-m)}{2^{m+1}}$$

的准确性(n 比特长序列 m 比特长模板)、计算理论方差的最优无偏估计

$$S^2 = \frac{1}{N} \sum_{i=1}^N (X_i - a)^2$$

的收敛性以及近似正态分布的符合性.我们选取单个样本长度为 1 GB($n=2^{33}$),重叠模板长度为 $m=64$ 比特和 56 比特进行样本实验.

在前述计算平台上单个样本(1GB 数据)的搜索时间约 400 s 左右,这里我们给出部分结果进行说明,需要说明的是 1000 个样本相对于整个 1 GB 可能的空间是非常小的,但依然可以观测到明显的统计性质.如图 5 所示,实际样本均值在理论均值附近波动,说明理论均值的准确性,而样本标准差随着样本量的增加表现出收敛特性.在进行具体实验设计时,可采用 Bootstrap 等方法得到合理的方差估计.如图 6 所示,对于 1 GB 数据而言,64 比特的重叠碰

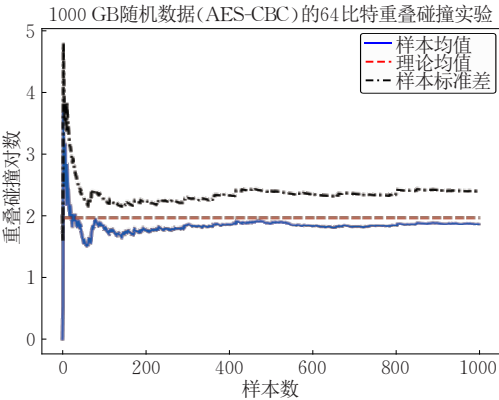


图 5 1GB 数据 64 比特重叠碰撞均值实验

撞对数存在明显的非对称性,这是由于 1 GB 数据极有可能不出现 64 比特碰撞,这些情形都统计为 0.而(图 7 所示)56 比特的重叠碰撞对数则表现出较好的近似正态性,应用我们提出的多样本 t 分布,得到 p 值为 0.024 436,通过检验.因此在设计具体检验工具时需特别注意近似符合性,选择合适的参数.图 8 给出了 1000 个 1 GB 样本数据的最长重叠碰撞长度分布情况,集中于 62~66 比特长度,最长碰撞长度为 75 比特,这与我们前面给出的 1 GB 的重叠碰撞拒绝检验策略考虑基本一致.

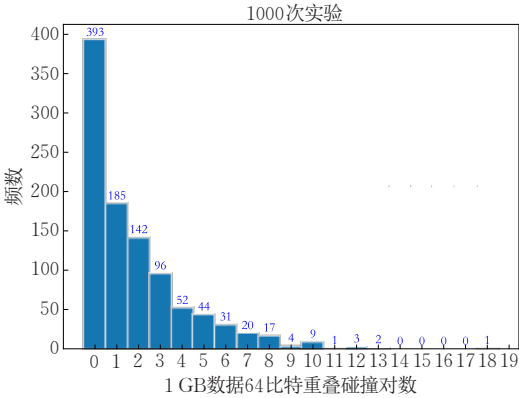


图 6 1GB 数据 64 比特重叠碰撞对数的分布实验

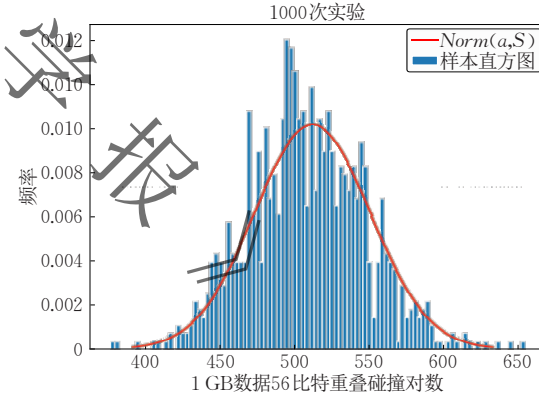


图 7 1GB 数据 56 比特重叠碰撞对数的分布实验

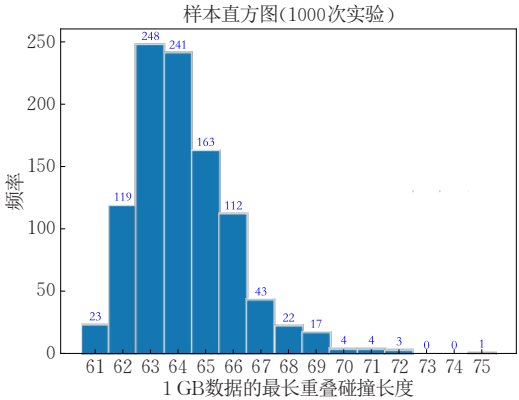


图 8 1GB 数据最长重叠碰撞长度分布实验

7 总 结

近年来信息处理和计算能力都有了较大的提升,而现有随机数统计检验一般规模参数较小,或者仅仅是大样本量的多次小样本重复检验,同时很多检验方法在大样本量可靠性存在问题,因此需要研究大样本条件下的高效检验方法,而本文关注的序列重叠碰撞问题便是一个研究方向. 重叠碰撞问题中存在很多复杂的依赖关系,这导致其统计特征难以有效估计,本文引出了重叠碰撞对概念,以序列重叠碰撞对的个数为统计量,给出了其理论期望模型,得到了准确的 n 比特长序列 m 比特模板的重叠碰撞对数的理论期望值,在此基础上提出了两种高效的重叠碰撞检验方法,其检验规模(模板长度)达到了 64 比特以上. 同时,为了保证检验的高效性和实际可操作性,我们提出了资源受限下基于 CUDA 的快速搜索方法,通过分类、排序等技术手段,降低资源消耗、提高算法并行性,从而能够高效地完成重叠碰撞搜索. 注意到,重叠碰撞的方差问题依然有待研究,但我们给出了理论近似和样本估计两个方向,方差的估计是很必要的,因为在实际测评中,我们需要及时判定不随机的样本,其有助于我们进一步完善检验步骤、缩短判定时间、降低检验样本量(同等检验规模下). 在搜索工具方面,我们提出的优化方法可根据计算资源设定参数,而在具体的计算资源下依然有很多值得优化的地方,以期进一步提高效率. 这些工作我们将在后续的规范测评流程中继续完善研究.

参 考 文 献

[1] DSA-1571-1 openssl—predictable random number generator. Debian Security Advisory. <http://www.debian.org/security/2008/dsa-1571>, 2008

[2] Zhang Wen-Tao. The trust crisis between American mathematicians and NSA. World Science, <https://worldscience.cn/qk/2015/4y/dsjysztx/583984.shtml>, 2015(in Chinese)
(张文韬. 美国数学家与 NSA 之间的信任危机. 世界科学, 2015)

[3] L’Ecuyer P, Simard R. TestU01: A C library for empirical testing of random number generators. AMC Transactions on Mathematical Software, 2007, 33(4): 1-40

[4] Rukhin A, et al. A Statistical Test Suite for Random and Pseudorandom Number Generators for Cryptographic Applications. United States: National Institute of Standards and Technology (NIST) Special Publication 800-22, Rev.1a, Apr. , 2010

[5] Marsaglia G. The Marsagila random number CD-ROM including the DieHard battery of test of randomness. <http://stat.fsu.edu/pub/diehard/>, 1995

[6] Functionality Classes and Evaluation Methodology for Physical Random Number Generators, Version 2, English Translation. Germany: BSI, Application Notes and Interpretation of the Scheme (AIS) 31, 2011

[7] GM/T 0005. Random test specification. National Cryptography Administration, 2021(in Chinese)
(GM/T 0005. 随机性检验规范. 国家密码管理局, 2021)

[8] Sulak F. Statistical Analysis of Block Ciphers and Hash Functions [Ph. D. dissertation]. In partial fulfillment of the requirements for the degree of Philosophy of Doctorate in Department of Cryptography Middle East Technical University. Ankara, Turkey, 2011

[9] Coron J-S, Naccache D. An accurate evaluation of Maurer’s Universal Test//Proceedings of the SAC 98 (Lecture Notes in Computer Science). Berlin, Germany, 1998: 57-71

[10] Hamano K, Kaneko T. Correction of overlapping template matching test included in NIST randomness test suite. IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, 2007, 90-A(9): 1788-1792

[11] Pareschi F, Rovatti R, Setti G. On statistical tests for randomness included in the NIST SP800-22 test suite and based on the binomial distribution. IEEE Transactions on Information Forensics and Security, 2012, 7(2): 491-505

[12] Chen Meihui, Chen Hua, Fan Limin, et al. A new discrete Fourier transform randomness test. Science China Information Sciences, 2019, 62(3): 1-16

[13] Barbour A D, Holst L, Janson S. Poisson Approximation. United Kingdom: Oxford University Press (Section 8. 4 and Section 10. 4), 1992

[14] Chrysaphinou O, Papastavridis S. A limit theorem on the number of overlapping appearances of a pattern in a sequence of independent trials. Probability Theory and Related Fields, 1988, 79(1): 129-143

[15] Rukhin A. Approximate entropy for testing randomness. Journal of Applied Probability, 2000, 37(1): 88-100

[16] Good I J. The serial test for sampling numbers and other tests for randomness. Mathematical Proceedings of Cambridge Philosophical Society, 1953, 49(2): 276-284

[17] CUDA. <http://developer.nvidia.com/about-cuda>

[18] Radix sort. http://en.wikipedia.org/wiki/Radix_sort

[19] Fu J C, Johnson B C. Approximate probabilities for runs and patterns in I. I. D. and Markov-dependent multistate trials. Advanced in Applied Probability, 2009, 41(1): 292-308

[20] Martin D E K. Distributions of pattern statistics in sparse Markov models. *Annals of the Institute of Statistical Mathematics*, 2020, 72(4): 895-913

[21] Martin D E K. Computation of exact probabilities associated with overlapping pattern occurrences. *WIREs Computational Statistics*, 2019, 11(5): e1477

[22] Menezes A J, van Oorschot P C, Vanstone S A. *Handbook of Applied Cryptography*. USA: CRC Press, 1997

[23] Casella G, Berger R L. *Statistical Inference (2nd Edition)*. Singapore: Cengage Learning, 2001

附录 A. 10 GB 数据的重叠碰撞搜索实例.

依据所用计算资源以及重叠碰撞理论期望模型,具体 10 GB 数据的重叠碰撞搜索步骤如下所示:

- S1. 初始参数设置. 设定检查模板长度 14(字节)、前缀长度 2(字节)、后缀长度 12(字节). 设定 CUDA 并行线程: 1024 个匹配线程块,每块含 32 个线程; 256 个排序线程块,每块含 256 个线程. 设定输出碰撞模板长度范围: 64 ~ 112 比特.
- S2. 数据载入. 将 10 GB 数据复制到 GPU 设备.
- S3. 匹配统计. 对数据进行 CUDA 并行匹配(《1024, 32》),

- 得到 2^{16} 个前缀的频次.
- S4. 匹配排序. 遍历第一个前缀字节(共 256 个),CUDA 并行排序(《256, 256》),每个线程块针对一个第二前缀字节(共 256 个),用 256 个线程对后缀 12 字节数据集进行并行基数排序(具体流程为:首先统计后缀数据的第一字节的偏移情况,然后每个线程针对后缀数据后 11 字节串行实现基数排序).
- S5. 输出. 将结果传输到 CPU 端进行结果处理并输出 {碰撞长度;碰撞模板;碰撞位置}.



CHEN Dong-Yu, Ph. D. candidate. His main research interests include the design and analysis of random number generator.

FAN Li-Min, Ph. D. , senior engineer. Her main research interests include side channel analysis and protection, cryptog-

raphy detection.

CHEN Hua, Ph. D. , professor-level senior engineer. Her main research interests include side channel analysis and protection, cryptography detection.

WANG Jian, Ph. D. candidate. His main research interests include side channel analysis and protection.

BU Yi-Fang, M. S. candidate. His main research interests include the design and analysis of random number generator.

Background

In recent years, the information processing and computing capabilities have been greatly improved. However, the scale parameters of the existing classical statistical tests for randomness are generally small, or they are just repeated tests with large sample size for many times with small samples. At the same time, there are problems in the reliability of many tests with large samples. Therefore, it is necessary to develop efficient and high-scale test methods, and the sequence overlapping collision problem that this paper focuses on is one of the research directions. There are many complex dependencies in the overlapping collision problem, which makes it difficult to estimate its statistical characteristics effectively. In this paper, we introduce the concept of overlapping collision pairs. Taking the number of sequence overlapping collision pairs as

the statistic, the theoretical expectation model is given, that is, the accurate theoretical expectation value of the statistic with n -bit sequence and m -bit template is obtained. On this basis, we propose two efficient overlapping collision tests, whose scale parameter (template length) reaches more than 64 bits. At the same time, in order to ensure the efficiency and practical operability of the test, we have developed a rapid search method based on CUDA under resource constraints, that is, with the skills such as sorting by classification, we can reduce the resource consumption and improve the parallelism of the algorithm, so that we can effectively complete the overlapping collision search.

This work is supported by the National Key Research of China (No. 2020YFA0309704).