Random Subsets of the Interval and P2P Protocols ⋆

Jacek Cichoń, Marek Klonowski, Łukasz Krzywiecki, Bartłomiej Różański, and Paweł Zieliński

Institute of Mathematics and Computer Science
Wrocław University of Technology
Poland
{Jacek.Cichon, Marek.Klonowski, Lukasz.Krzywiecki,
Pawel.Zielinski}@pwr.wroc.pl

Abstract. In this paper we compare two methods for generating finite families of random subsets according to some sequence of independent random variables $\zeta_1, \ldots, \zeta_n$ distributed uniformly over the interval $[0, 1]$. The first method called *uniform split* uses ζ_i values straightforwardly to determine points of division of $[0, 1]$ into subintervals. The second method called *binary split* uses ζ_i only to perform subsequent divisions of already existing subintervals into exact halves. We show that the variance of lengthes of obtained intervals in the first method is approximately $\frac{1}{n^2}$ and that the variance of lengthes of obtained intervals in the second method is approximately $\frac{1}{n^2}(\frac{1}{\ln 2} - 1)$.
The uniform split is used in the Chord peer-to-peer protocol while the binary split is used in the CAN protocol. Therefore our analysis applies to this protocols and shows that CAN has a better probabilistic properties than Chord. We propose also a simple modification of the Chord protocol which improves its statistical properties.

1 Introduction

We investigate the problem of splitting a given interval into a finite number of nonoverlapping subintervals that appears in some peer-to-peer protocols. Splitting is done according to a sequence of random values $\zeta_1, \ldots, \zeta_n$ distributed uniformly in $[0, 1]$, and some fixed split method.

In this paper we present an analysis of two split methods. The first among them is rather straightforward. The family of subintervals is composed of all nonoverlapping intervals defined by the set $\{0, 1\} \cup \{\zeta_1, \ldots, \zeta_n\}$. This method corresponds to the sequential splitting by adding points – for a new point ζ_i we select an interval $(\zeta_j, \zeta_k]$ such that $\zeta_j < \zeta_i \leq \zeta_k$ and divide it into two parts: $(\zeta_j, \zeta_i]$ and $(\zeta_i, \zeta_k]$. We call this method *uniform split*. It is well known that this method has significant flaws in terms of subset length uniformity. Note that the

⋆ Partially supported by the EU within the 6th Framework Programme under contract 001907 (DELIS).

uniform split corresponds to the process of adding a nodes in the Chord peer-to-peer protocol (see [1] and [2]). It is well known, and our calculation confirms this fact, that the capacity of Chord nodes' areas (intervals) is a random variable with large variation. Now let us recall that amount of data and the number of requests passed via node in Chord is proportional to the length of its area. Hence, large variation of area size introduces a discrepancy between nodes' workload.

The second method called *binary split* is based on the following idea: if ζ_i values are used only to determine which existing interval is to be split, the splitting point is chosen always in the middle of selected interval. We show in Section 2 that the uniformity of interval lengths is significantly better than in the uniform split case. Binary split corresponds to a sequential process where each ζ_i determines an existing interval to split in two halves. The process starts with whole interval $[0, 1]$. ζ_1 obviously splits it into $[0, 0.5]$ and $(0.5, 1]$, but ζ_2 may either make it $[0, 0.25]$, $(0.25, 0.5]$ and $(0.5, 1]$ or $[0, 0.5]$, $(0.5, 0.75]$ and $(0.75, 1]$– depending on which initial interval ζ_2 falls in; and so on for all ζ_i for $1 < i \leq n$. The resulting family consists of $n+1$ nonoverlapping intervals with lengths from the set $\{\frac{1}{2^k} : k \leq n+1\}$. The binary split corresponds to the process of adding nodes in CAN peer-to-peer protocol (see [3]). In the last section of this paper we shall propose a small modification of the classical Chord protocol which is based on the binary split and which has better probabilistic properties than the original one.

The authors wish to express thanks to referees for their helpful suggestions concerning the presentation of this paper.

1.1 Notation

We denote the real numbers and integers by $\mathbb{R}$ and $\mathbb{Z}$, respectively. Let X be a random variable. We denote its expected value, variance and standard deviation by $\mathbf{E}[X]$, $\mathbf{var}[X]$ and $\mathbf{std}[X]$, respectively.

Let f be a complex function. We denote the residuum of the function $f(z)$ at the point a by $\mathrm{Res}\,[f(z)|z = a]$ (see [4]). The imaginary unit is denoted by $\mathbf{i}$, the real and imaginary parts of the complex number z are denoted by $\Re(z)$ and $\Im(z)$, respectively.

1.2 Arbitrary Split Method

Let $\mathcal{P}_n$ be any randomized method of generating a random set of n points from the interval $[0, 1]$. The set $\mathcal{P}_n(\omega)$ defines a sequence $(x_1^{\mathcal{P}_n(\omega)}, \ldots, x_{n+1}^{\mathcal{P}_n(\omega)})$ of lengths of consecutive intervals. By definition $x_1^{\mathcal{P}_n(\omega)} + \ldots + x_{n+1}^{\mathcal{P}_n(\omega)} = 1$, hence $\frac{1}{n+1}(x_1^{\mathcal{P}_n(\omega)} + \ldots + x_{n+1}^{\mathcal{P}_n(\omega)}) = \frac{1}{n+1}$. Let

$$\mathbf{var}[\mathcal{P}_n] = \mathbf{E}\left[\frac{1}{n+1}\sum_{i=1}^{n+1}(x_i^{\mathcal{P}_n} - \frac{1}{n+1})^2\right]$$

and $\mathbf{std}[\mathcal{P}_n] = \sqrt{\mathbf{var}[\mathcal{P}_n]}$. We may treat the number $\mathbf{std}[\mathcal{P}_n]$ as a measure of non-uniformity of distribution of points from a random set of cardinality n generated by process $\mathcal{P}_n$. It is easy to check that

$$\mathbf{var}[\mathcal{P}] = \frac{1}{n+1}\left(\mathbf{E}\left[\sum_{i=1}^{n+1}(x_i^{\mathcal{P}_n})^2\right] - \frac{1}{n+1}\right) . \tag{1}$$

Let us fix some subset $a = \{a_1, \ldots, a_n\}$ of $[0,1]$, and let us now choose some random point $\zeta \in [0,1]$ according to the uniform distribution in $[0,1]$. Then there exists an unique subinterval I generated by points of a such that $\zeta \in I$. We call this interval a *randomly uniformly chosen interval.* Let us recall the following basic fact:

Theorem 1. *Let $\mathcal{P}_n$ be an arbitrary method of generation of a random subset $\{a_1, \ldots, a_n\}$ of the interval $[0,1]$. Then the number*

$$\mathbf{ELRI}[\mathcal{P}_n] = \mathbf{E}\left[\sum_{i=1}^{n+1}(x_i^{\mathcal{P}})^2\right] \tag{2}$$

is the expected value of the length of randomly chosen interval.

Proof. Let $I_1(\omega), \ldots I_{n+1}(\omega)$ be the sequence of intervals generated by the set $\mathcal{P}(\omega)$. Let ζ be a random number from the uniform distribution on $[0,1]$ and let $L(\omega,\zeta)$ be the length of this interval $I_i(\omega)$ that $\zeta \in I_i(\omega)$. Then we have

$$\mathbf{ELRI}[\mathcal{P}_n] = \int_{\Omega\times[0,1]} L(\omega,x)(dP\times d\lambda)(\omega,x) =$$

$$= \int_{\Omega}\left(\int_0^1 L(\omega,x)dx\right)dP(\omega) = \int_{\Omega}\left(\sum_{i=1}^{n+1}|I_i(\omega)|\cdot \Pr(x\in I_i(\omega))\right)dP(\omega) =$$

$$\int_{\Omega}\left(\sum_{i=1}^{n+1}|I_i(\omega)|^2\right)dP(\omega) = \mathbf{E}\left[\sum_{i=0}^{n+1}(x_i^{\mathcal{P}})^2\right] .$$

□

1.3 The Uniform Split

Let us consider a sequence $X_1, \ldots, X_n$ of independent uniformly distributed in $[0,1]$ random variables. They generate a random subset $\{X_1, \ldots, X_n\}$ of the interval $[0,1]$ and we denote this method by $unif_n$ and call it *uniform split* (see [5]). The set $\{X_1, \ldots, X_n\}$ induces a partition of $[0,1]$ into n subintervals whose lengths, taken in proper order, will be denoted by $x_1, \ldots, x_{n+1}$. Then for any $t_1 \geq 0, \ldots, t_{n+1} \geq 0$ we have

$$\Pr(x_1 \geq t_1, \ldots, x_{n+1} \geq t_{n+1}) = (1-(t_1+\ldots+t_{n+1}))_+^n , \tag{3}$$

where $(a)_+ = \max\{a,0\}$ (see Feller [6]).

Let us consider the random variable x_1, i.e. the length of the first interval. From equation (3) we see that $\Pr(x_1 \geq t) = (1-t)^n$ for $t \in [0,1]$. Therefore the density of the variable x_1 equals $\varphi(t) = (1-(1-t)^n)' = n(1-t)^{n-1}$. Notice also that the remaining variables $x_2, \ldots, x_{n+1}$ have the same density as x_1.

Theorem 2. $\mathbf{ELRI}[unif_n] = \frac{2}{n+2}$

Proof. The result follows from the following direct calculations:

$$\mathbf{ELRI}[unif_n] = \mathbf{E}\left[\sum_{i=1}^{n+1} x_i^2\right] = (n+1)\mathbf{E}\left[x_1^2\right] = (n+1)\int_0^1 x^2\varphi(x)dx =$$

$$(n+1)n\int_0^1 x^2(1-x)^{n-1}dx = \frac{2}{n+2} \; .$$

□

Remark 1. We used the identity $\int_0^1 x^2(1-x)^{n-1}dx = \frac{2}{n(n+1)(n+2)}$ which can be proved by induction on n or can be evaluated by the use of the Euler beta function $\int_0^1 x^{a-1}(1-x)^{b-1}dx = \frac{\Gamma(b)\Gamma(b)}{\Gamma(a+b)}$.

From Theorem 1, Theorem 2 and Equation (1) we get:

Corollary 1. $\mathbf{var}[unif_n] = \frac{n}{(1+n)^2(2+n)}$.

2 The Binary Split

Let us fix a natural number n. Let us consider the following method of generation of a random subset of $[0,1]$ of cardinality n. We start from an empty set of points. Suppose we already have set A_k of points from $[0,1]$ and a new point a_{k+1} is to be added. We choose a random point $y \in [0,1]$ and select an interval I generated by points from the set A_k such that $y \in I$. Then we define a_{k+1} as the the middle point of the interval I and put $A_{k+1} = A_k \cup \{a_{k+1}\}$. We stop this process after n steps. We call this method the *binary split* and we denote this method by bin_n.

Our goal is to calculate the value of $\mathbf{var}[bin_n]$. Let us start with putting $f_0 = 1$ and $f_n = \mathbf{ELRI}[bin_n]$ for $n > 0$.

Lemma 1. *For all* $n \in \mathbb{N}$ *we have*

$$f_{n+1} = \frac{1}{2^{n+1}} \sum_{k=0}^{n} \binom{n}{k} f_k \; . \tag{4}$$

Proof. Let us consider the sequence $(\xi_1, \ldots, \xi_n, \xi_{n+1})$ of independent random variables uniformly distributed in the interval $[0,1]$ defined on a probabilistic space Ω and let $\omega \in \Omega$. At the beginning the number $\xi_1(\omega)$ splits $[0,1]$ into two equal parts: $[0, 0.5]$ and $(0.5, 1]$. Let us now define $A = \{i > 1 : \xi_i(\omega) \leq 0.5\}$ and $B = \{i > 1 : \xi_i(\omega) > 0.5\}$. Then the variables $\{\xi_i : i \in A\}$ can only split the $[0, 0.5]$ interval while variables from the set $\{\xi_i : i \in B\}$ can only split the $(0.5, 1]$. Note that $\{2\xi_i : i \in A\}$ split the interval $[0, 1]$, hence $\mathbf{E}[(2\xi_i)_{i\in A}] = f_{|A|}$. A similar observation is true for the sequence $(2\xi_i - 1)_{i\in B}$. Therefore we have

$$f_{n+1} = \sum_{A\subseteq\{2,\ldots,n+1\}} \left(\frac{1}{4} f_{|A|} + \frac{1}{4} f_{|B|}\right) \left(\frac{1}{2}\right)^n =$$

$$= \frac{1}{2^{n+2}} \sum_{k=0}^{n} \binom{n}{k} (f_k + f_{n-k}) = \frac{1}{2^{n+1}} \sum_{k=0}^{n} \binom{n}{k} f_k \ .$$

□

Let

$$L_n = \prod_{j=n}^{\infty} (1 - \frac{1}{2^j}) \ .$$

It is easy to calculate that $L_1 \simeq 0.2888$ and easy estimations shows that the inequalities $1 - \frac{4}{2^n} < L_n < 1 - \frac{1}{2^n}$ holds for each $n \geq 1$. We shall express numbers f_n in terms of numbers L_n.

Lemma 2. $f_n = \sum_{m\geq 0} (\frac{1}{2})^m (1 - (\frac{1}{2})^m)^n L_{m+1} \quad .$

Proof. Let us consider the exponential generating function

$$x(t) = \sum_{n\geq 0} f_n \frac{t^n}{n!}$$

of the sequence $(f_n)_{n\geq 0}$. From equation (4) we get

$$x'(t) = \sum_{n\geq 0} f_{n+1} \frac{t^n}{n!} = \sum_{n\geq 0} \frac{1}{2^{n+1}} \sum_{k=0}^{n} \binom{n}{k} f_k \frac{t^n}{n!} = \tag{5}$$

$$\frac{1}{2} \sum_{n\geq 0} \left(\sum_{k=0}^{n} \binom{n}{k} f_k \right) \frac{(t/2)^n}{n!} \ , \tag{6}$$

hence the function $x(t)$ satisfies the following functional equation

$$2x'(2t) = x(t)e^t \ ,$$

i.e. $x'(t) = \frac{1}{2} x(\frac{t}{2}) e^{\frac{t}{2}}$. If we put $X(t) = x(t)e^{-t}$ then we obtain a slightly simpler equation

$$X'(t) = \frac{1}{2} X\left(\frac{t}{2}\right) - X(t) \ .$$

which can be solved explicitly. Namely we have

$$X(t) = \sum_{n\geq 0} \frac{t^n}{n!} (-1)^n \prod_{k=1}^{n} \left(1 - \left(\frac{1}{2}\right)^k\right) \ .$$

Since $x(t) = X(t)e^t$ we obtain

$$f_n = \sum_{k=0}^{n} \binom{n}{k} (-1)^k \prod_{j=1}^{k} \left(1 - \left(\frac{1}{2}\right)^j\right) \ .$$

The above formula is hard to be calculated accurately because it contains large coefficients with alternating signs. Therefore we need to transform it into a more suitable form. We put into the Euler partition formula (see [7])

$$\prod_{k=0}^{\infty} \frac{1}{1 - q^k z} = \sum_{n \geq 0} \frac{z^n}{\prod_{k=1}^{n}(1 - q^k)}$$

values $z = q^{a+1}$ and $q = \frac{1}{2}$, and get

$$\frac{1}{\prod_{k=a+1}^{\infty}(1 - (\frac{1}{2})^k)} = \sum_{n \geq 0} \frac{(\frac{1}{2})^{(a+1)n}}{\prod_{k=1}^{n}(1 - (\frac{1}{2})^k)} \ .$$

After multiplying both sides of this equality by L_1 we get

$$\prod_{j=1}^{a} \left(1 - \frac{1}{2^j}\right) = \sum_{n \geq 0} \left(\frac{1}{2}\right)^{(a+1)n} L_{n+1} \ .$$

and hence we obtain

$$f_n = \sum_{k=0}^{n} \binom{n}{k} (-1)^k \sum_{m \geq 0} \left(\frac{1}{2}\right)^{(k+1)m} L_{m+1} =$$

$$= \sum_{m \geq 0} L_{m+1} \left(\frac{1}{2}\right)^m \sum_{k=0}^{n} \binom{n}{k} (-1)^k \left(\left(\frac{1}{2}\right)^m\right)^k =$$

$$= \sum_{m \geq 0} \left(\frac{1}{2}\right)^m \left(1 - \left(\frac{1}{2}\right)^m\right)^n L_{m+1} \ ,$$

which proves the lemma. □

Remark 2. From Lemma 2 we may deduce that $f_{n+1} < f_n$ for each n.

Let us consider now the following function

$$\varphi_n(x) = \left(\frac{1}{2}\right)^x \left(1 - \left(\frac{1}{2}\right)^x\right)^n$$

defined on the interval $[0, \infty)$. The function φ_n has the global maximum at point $\log_2(n+1)$ and $\varphi_n(\log_2(n+1)) = \frac{1}{ne} + o(\frac{1}{n})$. Notice that $\sum_{m \geq 0}(\frac{1}{2})^m(1-(\frac{1}{2})^m)^n = \sum_{m \geq 0} \varphi_n(m)$. Moreover, $\int_0^{\infty} \varphi_n(x)dx = \frac{1}{(1+n)\ln 2}$. From these observations we deduce that

$$\sum_{m \geq 0} (\frac{1}{2})^m (1 - (\frac{1}{2})^m)^n = O(\frac{1}{n}) \ . \qquad (7)$$

Lemma 3. $f_n = \sum_{m\geq 0} \frac{1}{2^m}(1-\frac{1}{2^m})^n + o(\frac{1}{n})$.

Proof. The proof is done by a simple estimation. Let us first show the following approximation

$$\sum_{m=0}^{\log_2 \sqrt{n}} \frac{1}{2^m}(1-\frac{1}{2^m})^n = o(\frac{1}{n}) \ .$$

This fact follows immediately from monotonicity of the function φ_n on the interval $[0, \log(n+1)]$. Namely,

$$\varphi_n(\log_2 \sqrt{n}) < \frac{1}{\sqrt{n}e^{\sqrt{n}}} \ ,$$

so

$$\sum_{m=0}^{\log_2 \sqrt{n}} \frac{1}{2^m}(1-\frac{1}{2^m})^n \leq \frac{\log_2 \sqrt{n}}{\sqrt{n}e^{\sqrt{n}}} \leq \frac{1}{e^{\sqrt{n}}}$$

and $\frac{1}{e^{\sqrt{n}}} = o(\frac{1}{n})$.

Observe that if $k > \log_2 \sqrt{n}$ then $L_k > 1 - \frac{4}{\sqrt{n}}$, so we have

$$(1-\frac{4}{\sqrt{n}}) \sum_{m>\log_2 \sqrt{n}} \frac{1}{2^m}(1-\frac{1}{2^m})^n \leq \sum_{m>\log_2 \sqrt{n}} \frac{1}{2^m}(1-\frac{1}{2^m})^n L_{m+1}$$

and from equation (7) we obtain

$$\left| \sum_{m>\log_2 \sqrt{n}} \frac{1}{2^m}(1-\frac{1}{2^m})^n - \sum_{m>\log_2 \sqrt{n}} \frac{1}{2^m}(1-\frac{1}{2^m})^n L_{m+1} \right| \leq \frac{C}{n\sqrt{n}} = o(\frac{1}{n}) \ ,$$

which proves the lemma. □

Lemma 4. $\sum_{k\geq 0} \frac{1}{2^k}(1-\frac{1}{2^k})^n = \sum_{k=0}^n \binom{n}{k}(-1)^k \frac{1}{1-(\frac{1}{2})^{1+k}}$

Proof. The proof follows from following transformations:

$$\sum_{k\geq 0} \frac{1}{2^k}(1-\frac{1}{2^k})^n = \sum_{k\geq 0}(\frac{1}{2^k}) \sum_{l=0}^{n} \binom{n}{l}(-1)^l \frac{1}{2^{kl}} =$$

$$\sum_{l=0}^{n} \binom{n}{l}(-1)^l \sum_{k\geq 0}(\frac{1}{2})^{kl+k} = \sum_{l=0}^{n} \binom{n}{l}(-1)^l \frac{1}{1-(\frac{1}{2})^{1+l}} \ .$$

□

Theorem 3. *The sequence f_n satisfies*

$$f_n = \frac{1}{n+1}\left(\frac{1}{\ln 2} + \omega(\log_2(n+1)) + \eta(n)\right) + O(\frac{1}{n^2}) \ ,$$

where ω is a periodic function with period 1 such that $|\omega(x)| < 1.42602 \cdot 10^{-5}$ and $|\eta(x)| < 6.72 \cdot 10^{-11}$.

In the proof of this theorem we use a method of the treatment of oscillating sums attributed to S.O. Rice by D.E. Knuth (see [7]).

Proof. Let us, for simplicity, denote

$$s_n = \sum_{k=0}^{n} \binom{n}{k} (-1)^k \frac{1}{1-(\frac{1}{2})^{1+k}}$$

and let us consider the following function

$$f(z) = \frac{1}{1-(\frac{1}{2})^{1-z}}$$

with complex argument z. Then f is a meromorphic function with single poles at points

$$\mathfrak{z}_k = 1 + \frac{2\pi k}{\ln 2}\mathbf{i} \ ,$$

where $k \in \mathbb{Z}$, $\mathbf{i}$ is the imaginary unit and

$$s_n = \sum_{k=0}^{n} \binom{n}{k} (-1)^k f(-k) \ .$$

Let $B(x, y) = \Gamma(x)\Gamma(y)/\Gamma(x + y)$ be the Euler beta function. The function $B(n + 1, z)$ has single poles at points $0, -1, \ldots, -n$ and

$$\mathrm{Res}\,[B(n+1,z)|z=-k] = (-1)^k \binom{n}{k}.$$

(see [7] for details). Notice that the function f is holomorphic on the half-plane $\Re(z) < 0.5$. Therefore

$$s_n = \sum_{k=0}^{n} \mathrm{Res}\,[B(n+1,z)f(z)|z=-k] \ . \tag{8}$$

Let us consider a big rectangle C_k with end-point $\pm k \pm (2k+1)\pi\mathbf{i}/\ln 2$, where k is a natural number. It is quite easy to check that

$$\lim_{k\to\infty} \oint_{C_k} B(n+1,z)f(z)dz = 0$$

(in proof of this fact the equality $f(1 + (2k + 1)\pi\mathbf{i}/\ln 2) = 1/2$ plays a crucial role). Therefore, using Cauchy Residue Theorem (see [4]), we get

$$s_n = -\sum \{\mathrm{Res}\,[B(n+1,z)f(z)|z=\mathfrak{z}_k] : k \in \mathbb{Z}\} \ . \tag{9}$$

Further, it can be easily checked that

$$\mathrm{Res}\,[B(n+1,z)f(z)|z=1] = -\frac{1}{(n+1)\ln 2} \ . \tag{10}$$

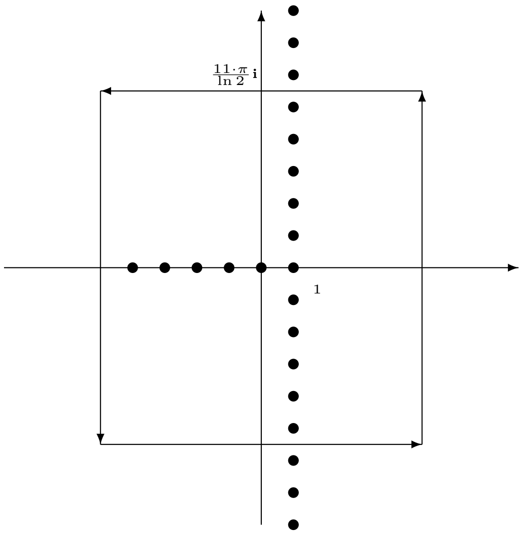

Fig. 1. Singular points of the function $B(4,z)f(z)$ and the contour of integration C_5

This first residue gives us the first part of the approximation of the number s_n. Generally, we have

$$\mathrm{Res}\,[B(n+1,z)f(z)|z=\mathfrak{z}_k] = -\frac{\Gamma(1+n)\Gamma(1+\frac{2k\pi\mathbf{i}}{\ln 2})}{\Gamma(2+n+\frac{2k\pi\mathbf{i}}{\ln 2})\ln 2}\ .$$

Notice that if $x\in\mathbb{R}$ then

$$\left|\frac{\Gamma(n+1)\Gamma(1+\mathbf{i}x)}{\Gamma(2+n+\mathbf{i}x)}\right| = \frac{n!}{\sqrt{(1^2+x^2)(2^2+x^2)\cdots((n+1)^2+x^2)}} \leq$$

$$\frac{1}{n+1}\cdot\frac{1}{\sqrt{(1+\frac{x^2}{1^2})(1+\frac{x^2}{2^2})\cdots(1+\frac{x^2}{n^2})}}\ .$$

Let $a=\frac{2\pi}{\ln 2}$. It can be checked numerically that

$$\sum_{k=2}^{\infty}\frac{1}{\sqrt{\prod_{m=1}^{100}(1+(ka/m)^2)}} \approx 2.32781\cdot 10^{-11}\ .$$

Therefore

$$|\sum\{\mathrm{Res}\,[B(n+1,z)f(z)|z=\mathfrak{z}_k] : |k|\geq 2\}| \leq \frac{6.72\cdot 10^{-11}}{n+1} \qquad (11)$$

for all $n\geq 100$. Next, we use the following well known approximation formula

$$\frac{\Gamma(z+a)}{\Gamma(z+b)} = z^{a-b}\left(1+\frac{(a-b)(a+b-1)}{2z}+O(\frac{1}{z^2})\right)\ ,$$

which holds when $|z| \to \infty$ and $|\arg(z+a)| < \pi$ to expressions

$$\frac{\Gamma(1+n)}{\Gamma(1+n+1 \pm \frac{2\pi\mathbf{i}}{\ln 2})} ,$$

and we obtain

$$\frac{\Gamma(1+n)}{\Gamma(1+n+1 \pm \frac{2\pi\mathbf{i}}{\ln 2})} =$$

$$\frac{(2\pi^2 + \mathbf{i}\pi\ln 2 + (n+1)\ln^2 2)\left(\cos\frac{2\pi\mathbf{i}\ln(n+1)}{\ln 2} \mp \mathbf{i}\sin\frac{2\pi\mathbf{i}\ln(n+1)}{\ln 2}\right)}{(n+1)^2\ln^2 2} + O(\frac{1}{n^2}) .$$

After noticing that $\Gamma(1 \pm \frac{2\pi\mathbf{i}}{\ln 2}) \approx 3.1766 \cdot 10^{-6} \mp 3.7861 \cdot 10^{-6} \cdot \mathbf{i}$ and some simple calculations we finally get

$$\sum\{\mathrm{Res}\,[B(n+1,z)f(z)|z = \mathfrak{z}_k] : |k| = 1\} = \tag{12}$$

$$\frac{10^{-6}}{n+1}\left(a \cdot \cos(2\pi\log_2(n+1)) - b \cdot \sin(2\pi\log_2(n+1))\right) + O(\frac{1}{n^2}) ,$$

where $a \approx 9.166$ and $b \approx 10.924$. Putting together Equations (9), (10), (11) and (12) we obtain the thesis of the theorem. □

From the last Theorem and Equation (1) we obtain

Corollary 2.

$$\mathbf{var}[bin_n] = \frac{1}{(n+1)^2}\left(\frac{1}{\ln 2} - 1 + \omega(\log_2(n+1)) + \eta(n)\right) + O(\frac{1}{n^3}) ,$$

where ω and η are the function from Theorem 3.

Notice that $\frac{1}{\ln 2} - 1 \approx 0.4427$, therefore $\mathbf{std}[bin_n] \approx \frac{0.665}{n}$, hence $\mathbf{std}[bin_n]$ is significantly smaller than the value $\mathbf{std}[unif_n] \approx \frac{1}{n}$ (see Corollary 1). From this we conclude that the binary split process generates a more uniform distribution of random points in the interval $[0,1]$ than the uniform one.

Remark 3. The main influence on the asymptotic behavior of the sequence (f_n) is played by the three main poles of the function $f(z) = (1 - (\frac{1}{2})^{1-z})^{-1}$: $\mathfrak{z}_0 = 1$, $\mathfrak{z}_1 = 1 + \frac{2\pi\mathbf{i}}{\ln 2}$ and $\mathfrak{z}_{-1} = 1 - \frac{2\pi\mathbf{i}}{\ln 2}$. The first one, located at point 1, is responsible for the component $\frac{1}{(n+1)\ln 2}$. The next two poles are responsible for relatively small oscillations of the sequence f_n. The size of oscillations is relatively small because $\Im(1 + 2\pi\mathbf{i}/\ln 2) \approx 10$ and the function Γ decreases rapidly when the imaginary part of a number grows.

Remark 4. Some aspects of the binary split model, namely the properties of the length of the first node, were investigated P. Flajolet (see [8]) and later by P. Kirchenhofer and H. Prodinger (see [9]) in their analyses of properties of the R. Morris probabilistic counter (see [10]). In their investigations a fluctuation factor of the form $\omega(\log_2 n)$ appears, too.

3 Discussion

Let $\mathcal{P}_n$ be any randomized method of generating random subsets of interval $[0,1]$ of cardinality n. Let

$$\mathbf{CV}[\mathcal{P}_n] = \frac{\mathbf{std}[\mathcal{P}_n]}{\mathbf{E}\left[\mathcal{P}_n\right]}$$

denotes the coefficient of variation of $\mathcal{P}_n$. The number $\mathbf{CV}[\mathcal{P}_n]$ is a measure of dispersion of lengths of intervals generated by the method $\mathcal{P}_n$. We have proved that $\mathbf{CV}[unif_n] \simeq 1$ and $\mathbf{CV}[bin_n] \simeq 0.665$. Therefore the random subsets generated by the binary split methods has smaller dispersion than subsets generated by the uniform method.

The length of intervals in the uniform split with n elements varies between $\frac{1}{n^2}$ and $\frac{\ln(n)}{n}$. To be more precise let us consider the following two random variables $\min(\mathcal{P}_n) = \min\{x_i^{\mathcal{P}_n}, \ldots, x_{n+1}^{\mathcal{P}_n}\}$, $\max(\mathcal{P}_n) = \max\{x_i^{\mathcal{P}_n}, \ldots, x_{n+1}^{\mathcal{P}_n}\}$. It follows almost directly from Equation (3) that $\mathbf{E}\left[\min(unif_n)\right] = \frac{1}{n^2}$. Moreover $\mathbf{E}\left[\max(unif_n)\right] \sim \frac{\ln n}{n} + \frac{\gamma}{n}$ (see [11]). These observations were done by several authors working with P2P protocols–see e.g. [12].

Let $m_o = M_0 = 1$ and $m_n = \mathbf{E}\left[\min(bin_n)\right]$ and $M_n = \mathbf{E}\left[\max(bin_n)\right]$ for $n > 0$. Using similar arguments as in Lemma 1 we may show that

$$m_{n+1} = \frac{1}{2^{n+1}} \sum_{i=0}^{n} \binom{n}{i} \min(m_i, m_{n-i}) \ ,$$

$$M_{n+1} = \frac{1}{2^{n+1}} \sum_{i=0}^{n} \binom{n}{i} \max(M_i, M_{n-i}) \ .$$

Numerical calculations show that $\frac{0.4}{\ln\ln n}\frac{1}{n} \leq m_n < M_n \leq \frac{2.2 \ln\ln n}{n}$ for each $n \geq 5$, however we do not have a precise mathematical proof of this fact. This shows that the length of intervals in the binary split is much better concentrated near its medium value $\frac{1}{n}$ than in the uniform split.

4 Conclusions

Our mathematical analysis was motivated by the problems arising in in computer science. It is well known that the capacity of Chord (see [1] and [2]) nodes' areas (intervals) is a random variable with large variation. Our calculation of $\sigma(unif_n)$ confirms this fact. Now let us recall that amount of data and number of requests passed via node in Chord is proportional to the length of its area. Hence, large variation of area size introduces a discrepancy between nodes' workload.

This problem can be partially solved by a very simple modification of the original Chord protocol. The modification is based on *binary split* method which can be easily embedded into Chord's protocol. Namely, we need to modify only one of Chord's procedure, namely the procedure *join*. The original ,,join" procedure accepts a new node at an arbitrarily chosen position ζ_i. In the modification we can use ζ_i only to determine which interval the new node P will split upon arrival. Then the target interval is slitted into two halves and the node currently responsible for the interval will keep roughly half of the resources and P takes over the responsibility for the rest. In other words, instead of using random protocol address, the new node randomly and uniformly picks an interval and joins the Chord protocol precisely in the middle of this interval. The method described above can be treated as Chord protocol with a one dimensional CAN's split method (see [3]).

We propose the modification of the Chord protocol only in one point, namely in the ,,join" procedure. In reality Chord is a dynamic structure; nodes both leave and join the network. It is possible to modify the structure of remaining nodes after a single node leaves the system in such a way that after this modification we shall obtain a division generated by the binary split method, however, this is a quite complicated procedure. Our proposition is to ignore this fact.

We have made a lot of numerical experiments for checking what happens when we use the binary split only in the ,,join" procedure. Figure 2 contains a summary of one experiment. In this experiment we have build a Chord structure

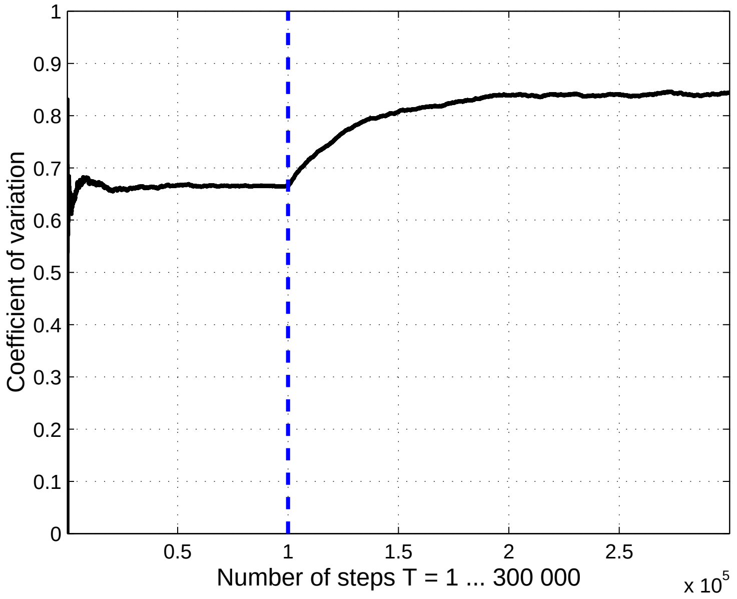

Fig. 2. Experiment with 10^5 nodes

based on the binary split method with 10^5 nodes and later we successively removed one randomly chosen node in the ,,normal way” and add one node using the binary split regime $2 \cdot 10^5$ times. We observed that afer the initialization phase the variable $\mathbf{CV}[\mathcal{P}]$ increase, but later its value stabilize and in the stable regime we have $\mathbf{CV}[\mathcal{P}] \approx 0.85$. This coefficient of variation is bigger than the coefficient of variation of the binary split but it is less than the coefficient of variation of the uniform split. However, this behavior requires and still awaits for precise theoretical explanation.

References

1. Stoica, I., Morris, R., Karger, D., Kaashoek, M.F., Balakrishnan, H.: Chord: A scalable peer-to-peer lookup service for internet applications. In: Proceedings of the ACM SIGCOMM ’01 Conference, San Diego, California, USA (2001)
2. Liben-Nowell, D., Balakrishnan, H., Karger, D.: Analysis of the evolution of peer-to-peer systems. In: 21st ACM Symposium on Principles of Distributed Computing (PODC), Monterey, CA (2002)
3. Ratnasamy, S., Francis, P., Handley, M., Karp, R., Shenker, S.: A scalable content-addressable network. In: Proceedings of the ACM SIGCOMM ’01 Conference, San Diego, California, USA (2001)
4. Cartan, H.: Elementary Theory of Analytic Functions of One or Several Complex Variables. Herman, Paris (1973)
5. Kopociński, B.: A random split of the interval $[0, 1]$. Aplicationes Mathematicae **31** (2004) 97–106
6. Feller, W.: An Introduction to Probability Theory and Its Applications. Volume II. John Wiley and Sons Inc, New York (1965)
7. Knuth, D.E.: Sorting and Searching. Third edn. Volume 3 of The art of computer programming. Addison-Wesley, Reading, Massachusetts (1997)
8. Flajolet, P.: Approximate counting: A detailed analysis. BIT **25** (1985) 113–134
9. Kirschenhofer, P., Prodinger, H.: Approximate counting: an alternative approach. Informatique Theorique et Applications **25** (1991) 43–48
10. Morris, R.: Counting large numbers of events in small registers. Communications of The ACM **21** (1978) 161–172
11. Devroye, L.: Laws of the iterated logarithm for order statistics of uniform spacings. The Annals of Probability **9**(5) (1981) 860–867
12. King, V., Saia, J.: Choosing a random peer. In: Proceedings of the 23rd Annual ACM Symposium on Principles of Distributed Computing. (2004) 125–130