THE PROBLEM OF ERROR FREQUENCY DISTRIBUTION IN THE MILLER-RABIN TEST FOR TRIPLEPRIME NUMBERS

Alisher Zhumaniezov

Kazan Federal University, Kazan, Russian Federation

ABSTRACT

This article investigates the error distribution of the Miller-Rabin test for the class of tripleprime numbers. At first the current results on the class of semiprimes are presented. Further, a theoretical estimation of the average frequency for triple prime numbers on an interval is derived, and a comparative analysis with a practical result is demonstrated. Graphs and intermediate conclusions accompany all comparisons. A conclusion is also made about a possible direction for improving this estimation.

KEYWORDS

Miller-Rabin test, strong pseudoprime, number theory, frequency distribution.

1. INTRODUCTION

Prime numbers are a key element in the design of cryptographic protocols. Therefore, it is very necessary to find a sufficiently large prime number quickly and efficiently. The easiest way is to iterate the numbers in a given interval and check for primality. Thus, the question arises of checking an arbitrary number for primality.

There are several approaches to checking whether a number is prime:

1. Search of divisors[1] – a deterministic algorithm that gives an answer in a limited time. The main disadvantage is a long running time, exponential dependence on the length of the number.

2. Fermat's test[2] – a probabilistic test based on Fermat's little theorem. The main disadvantage is the Carmichael numbers passing the test with an falsely primality verdict, and their infinite number.[3]

3. The Miller-Rabin test[4][5] – the most widely used probabilistic test. Is an improvement on the Fermat test.

There are many other primality tests, but they have not been considered in this paper.

The Miller-Rabin test is a probabilistic test, which means that this test may falsely conclude that a number is prime. Therefore, the problem arises of estimating the probability of such an error in order to evaluate the efficiency of the algorithm.

David C. Wyld et al. (Eds): FCST, CMIT, SE, SIPM, SAIM, SNLP - 2022
pp. 01-12, 2022. CS & IT - CSCP 2022 DOI: 10.5121/csit.2022.120801

For example, the article [6] provides an algorithm for additional verification of numbers that have passed the Miller-Rabin test with base 2. To correctly estimate its running time, it is necessary to know the distribution of strong pseudoprimes over this base.

Also, article [7] provides an algorithm for finding prime numbers by pattern. To estimate its running time, knowledge about the distribution of strong pseudoprimes was also needed.

In this paper, we present theoretical calculations for the average error of the Miller-Rabin test on the interval $[1, X]$. As part of the work, current estimation for numbers enclosed in parentheses and set on the right margin. For example, we present theoretical upper bound for semiprimes

$$n = pq \tag{1}$$

and calculate theoretical upper bound for tripleprimes

$$n = pqr \tag{2}$$

After the derivation of the upper bound, we perform practical calculations, according to which make conclusions about the closeness of the upper estimation to practical results and the need to improve the estimation.

This paper contains the following Sections: Section 2 present problem statement, all necessary definitions and current results of research. Sections 3 present obtaining of theoretical estimation for upper bound of average error on interval. Section 4 demonstrate visual comparison theoretical function and practical values on several parameters and made intermediate conclusions. Section 5 present final conclusion about the results.

2. METHODS

2.1. Miller-Rabin Test

There are many algorithms for checking the primality of a number. As part of this work, the Miller-Rabin probabilistic test will be analyzed. This test was developed in 1976 by G. Miller in the article [4]. Its modification was presented in 1980 by M. Rabin in the article [5]. This algorithm is based on Euler's theorem [8]:

$$a^{n-1} \equiv 1 (\mathrm{mod}\, n), \textit{where } n \textit{ is a prime number} \tag{3}$$

To describe the algorithm, we introduce the following definitions:

Definition 1. Let n be an arbitrary natural number, then we define the functions $bin(n)$ and $odd(n)$ as follows:

$$\begin{gathered} n = 2^s u, \textit{where } u \textit{ is odd} \\ bin(n) = s \\ odd(n) = u \end{gathered} \tag{4}$$

Definition 2. Let n is an arbitrary natural number, a is a natural number belonging to the interval $[1, n-1]$. Call a is witness of primality if one of the following conditions is met:

$$\begin{gathered} a^{odd(n-1)} \equiv 1 (\mathrm{mod}\, n) \\ \exists (0 \le i < bin(n-1)) \left| (a^{odd(n-1)})^{2^i} \equiv -1 (\mathrm{mod}\, n) \right. \end{gathered} \tag{5}$$

Thus, the final algorithm for checking for the primality of a number is as follows [5]:

Perform k iterations of the test:

Choose a random number a.

Find d = gcd(a,n). if d ≠ 1, then n is composite.

Check whether conditions from (5) are satisfied:

If none of the conditions is met, then n is a composite.

Otherwise - probably prime.

The verdict "probably prime" means that there is a composite number that will pass the test as a prime number. Such numbers are called strong pseudoprime. The study of the distribution of strong pseudoprimes is important for evaluating the efficiency of the algorithm.

One of the approaches to estimating the distribution of strictly pseudoprimes is sequence ψ_n – smallest strong pseudoprime to n first prime numbers as bases. Nowadays known values are for $1 \le n \le 13$ [9][10]. There is also conjecture for values $14 \le n \le 19$ [11]. However, in present work we use another approach.

2.2. Number of witnesses to the primality of the number

An important characteristic for estimating the error of the Miller-Rabin test is the number of witnesses to the primality of an arbitrary number. This value allows us to estimate the probability of choosing a witness of primality for a composite number, and, consequently, a falsely conclusion.

In the article [12], a necessary and sufficient condition was presented that allows finding the exact number of primality witnesses for an arbitrary number:

$$\begin{gathered} n = uv, where\, u\, and\, v\, are\, coprime \\ ord_u(a) \mid GCD(\varphi(u), (u - \varphi(u))v - 1) \\ ord_v(a) \mid GCD(\varphi(v), (v - \varphi(v))u - 1) \\ bin(ord_u(a)) = bin(ord_v(a)) \end{gathered} \tag{6}$$

We define by $W(n)$ the number of witnesses to the primality of n. The first formula for $W(n)$ was also presented in [12], but only for semiprime numbers n from (1).

$$W(n) = odd(d)^2 \frac{4^{bin(d)} + 2}{3}, where\, d = GCD(p-1, q-1) \tag{7}$$

In [13], a finite formula was presented for an arbitrary number by its decomposition into prime factors:

$$\begin{gathered} n = p_1^{r_1} * p_1^{r_1} * ... * p_k^{r_k} \\ d_i = GCD(p_i - 1, \frac{n}{p_i^{r_i}} - 1) \\ s = \min(bin(d_i)) \\ W(n) = \prod_{i=1}^{k} (odd(d_i)) * (1 + \sum_{j=0}^{s-1} 2^{kj}) \end{gathered} \tag{8}$$

2.3. Average frequency distribution

After introducing utility definitions and formulas, we define the function for calculations. Since in the Miller-Rabin test one of the checks is the calculation of the GCD of a number and base, then only numbers coprime with n can be witnesses of primality. Hence:

$$W(n) \le \varphi(n) \tag{9}$$

Thus, as the frequency of witnesses, we will define:

$$Fr(n) = \frac{W(n)}{\varphi(n)} \tag{10}$$

The maximum value 1 is reached if and only if n is prime. This follows from Rabin's theorem presented in [5]:

$$W(n) \le \frac{\varphi(n)}{4}, where\, n\, is\, a\, composite\, number \tag{11}$$

The value of the $Fr(n)$ function can also be interpreted as the probability of successfully passing one iteration of the Miller-Rabin test. In this case, the probability that the composite number n is probably prime after k iterations is $\frac{1}{4^k}$. Subsequently, this estimation was improved in [12] to $\frac{1}{16^k}$.

Since the function $Fr(n)$ has no limit and reaches a maximum of $\frac{1}{4}$ on an infinite number of composite numbers, we will estimate the distribution for $Avg(Fr(n))$ – the average frequency of witnesses on the interval $[1, X]$.

Article [14] presents estimation:

$$Avg(Fr(n)) < \frac{1}{\sqrt{X}} \tag{12}$$

Article [5] presents estimation:

1) for the case n from (1) and

$$\begin{gathered} q = (p-1)k+1 \\ Avg(Fr(n)) = \frac{p^2}{2X}\ln(\ln(X))\ln(X) \end{gathered} \tag{13}$$

2) for the case n from (1) and

$$\begin{gathered} q = 2k+1, where\ 2k \bmod (p-1) \neq 0 \\ Avg(Fr(n)) = \frac{2}{X}\ln(\ln(X))\ln(X) \end{gathered} \tag{14}$$

3) for the general case from (1)

$$E(Avg(Fr(n))) = \frac{2p}{X}\ln(\ln(X))\ln(X) \tag{15}$$

2.4. Information content

Since the average probabilities in practice are extremely small, we use the information content from [15] for the "probably prime" event to compare the theoretical and practical estimations:

$$I = \log(\frac{1}{p}), where\ p - event\ probability \tag{16}$$

3. RESULTS AND DISCUSSION

3.1. Estimating the number of witnesses

At first, we define d_1, d_2, d_3 and s for n from (2) where p, q, r are distinct prime numbers:

$$\begin{gathered} d_1 = GCD(p-1, qr-1) \\ d_2 = GCD(q-1, pr-1) \\ d_3 = GCD(r-1, pq-1) \\ s = \min(bin(d_1), bin(d_2), bin(d_3)) \end{gathered} \tag{17}$$

After that, we get the formula for the number of witnesses of the primality of number n from (2) where p, q, r are distinct prime numbers. Next, we calculate the inner sum in (8) through a geometric progression and get:

$$W(n) = odd(d_1) * odd(d_2) * odd(d_3) * \frac{8^s + 6}{7} \tag{18}$$

From the definition of s and the properties of the GCD it follows:

$$odd(d_1) = \frac{d_1}{2^{bin(d_1)}} \le \frac{d_1}{2^s} \le \frac{p-1}{2^s} \tag{19}$$

$$odd(d_2) = \frac{d_2}{2^{bin(d_2)}} \le \frac{d_2}{2^s} \le \frac{q-1}{2^s} \tag{20}$$

$$odd(d_3) = \frac{d_3}{2^{bin(d_3)}} \le \frac{d_3}{2^s} \le \frac{pq-1}{2^s} \tag{21}$$

But we can improve multiplication of the inequalities from (19), (20) and (21) by this theorem: For any d_1, d_2, d_3 and s, defined as in (17) this inequality is satisfied:

$$odd(d_1) * odd(d_2) * odd(d_3) \le \frac{(p-1)(q-1)(pq-1)}{2^{3s+1}} \tag{22}$$

Since the numbers $p-1$, $q-1$ and $pq-1$ are always even, then $s \ge 1$. Substituting this property and inequality from (22) into (18) we obtain:

$$W(n) \le \frac{(p-1)(q-1)(pq-1)(1+\frac{6}{8^1})}{14} \le \frac{(p-1)(q-1)(pq-1)}{8} \tag{23}$$

Thus, we obtain an estimation for the number of witnesses of the primality.

3.2. Estimating the frequency of witnesses

Substituting the inequality from (23) into (10) we get:

$$Fr(n) = \frac{W(n)}{\varphi(pqr)} \le \frac{(p-1)(q-1)(pq-1)}{8(p-1)(q-1)(r-1)} = \frac{pq-1}{8(r-1)} \tag{24}$$

Thus, we obtain an estimation for the frequency of witnesses to the primality of an arbitrary tripleprime number n from (2).

3.3. Estimating the average frequency of witnesses

To calculate the average frequency of witnesses, we fix prime numbers p and q and introduce the parameter y. We define by $S_{pq}(y)$ the sum of the frequencies of witnesses for all numbers n from (2), where r is a prime number on the interval $[\max(p,q)+1, y]$, and by $\pi_p(y)$ the number of prime numbers on the interval $[p+1, y]$.

Then we found the average frequency of witnesses on the segment $[1, pqy]$ by the formula:

$$Avg(Fr(n)) = \frac{S_{pq}(y)}{\pi_{\max(p,q)}(y)} \tag{25}$$

At first, we estimate $S_{pq}(y)$:

$$S_{pq}(y) \le \sum_{r \le y} \frac{pq-1}{8(r-1)} \approx \frac{pq-1}{8} \sum_{r \le y} \frac{1}{r} \approx \frac{pq-1}{8} \ln(\ln(y)) \tag{26}$$

After that, we estimate $\pi_p(y)$:

$$\pi_p(y) \approx \frac{y}{\ln(y)} - \frac{p}{\ln(p)} \approx \frac{y}{\ln(y)} \tag{27}$$

We denote the upper bound by X, then:

$$y \le \frac{X}{pq} \tag{28}$$

Substituting inequalities from (26), (28) and estimation from (27) into (25) we obtain an upper bound for $Avg(Fr(n))$. Then:

$$Avg(Fr(n)) \le \frac{(pq)^2 \ln(X) \ln(\ln(X))}{8X} \tag{29}$$

Thus, we obtained an estimation for the average frequency of witnesses to the primality of an arbitrary tripleprime number n from (2) on the interval $[1, X]$:

$$Avg(Fr(n)) \le C_{pq} \frac{\ln(X) \ln(\ln(X))}{X} \tag{30}$$

3.4. Convergence to upper bound

Since the inequalities in (19), (20) and (21) can be strengthened for an infinite number of numbers, the final upper bound is inaccurate and needs to be improved. However, we will try to consider the subproblem of finding $Avg'(Fr(n))$ – the average frequency of witnesses on the interval $[1, X]$ of tripleprimes number from (2), which satisfy the following conditions:

$$p - 1 = 2p', where\ p' - prime\ number \tag{31}$$

$$q - 1 = 2q', where\ q' - prime\ number \tag{32}$$

$$bin(pq-1) = 2 \tag{33}$$

$$\begin{aligned} rq &\equiv 1 (\mathrm{mod}\ p') \\ rp &\equiv 1 (\mathrm{mod}\ q') \\ r &\equiv 1 (\mathrm{mod}\ pq-1) \end{aligned} \tag{34}$$

There are infinitely many such triples p, q and r, since the equation (32) has exactly one solution r_0 on the segment $[1, p'*q'*(pq-1)]$. For example, if $p=7$, $q=11$ then $r_0=533$. Then all solutions of (32) will be $r=r_0+p'*q'*(pq-1)k$, where k – any integer. Thus, we can calculate $Avg'(Fr(n))$ over an arbitrarily large segment. In this case upper bound from (29) can be improved to close enough result:

$$Avg'(Fr(n)) \le \frac{pq(pq-1)}{8}\frac{\ln(X)(\ln(\ln(X))-\ln(\ln(y_0)))}{X} \qquad (35)$$

, where y_0 - minimum value r for fixed p and q. For example, if $p=7$, $q=11$ then $y_0=8513$.

4. SUMMARY

4.1. Comparison of theoretical and practical evaluation

After finding the theoretical estimates, we compare the result with practice for several points:

1. The value of the coefficient C_{pq} through the formula (30).
2. The ratio of both sides of the inequality in (29).
3. Comparison of both sides of the inequality in (29).
4. Convergence to upper bound from inequality in (35).
5. General conclusions.

4.2. Value of the coefficient C_{pq}

Figure 1 shows that the dependence of the coefficient is non-linear and not even monotonous. However, it can be noted that the type of dependence itself does not depend on the boundary of the segment, but only on pq. Also, it does not reach the theoretical value, so we can conclude that it is possible to improve the value of the coefficient C_{pq}.

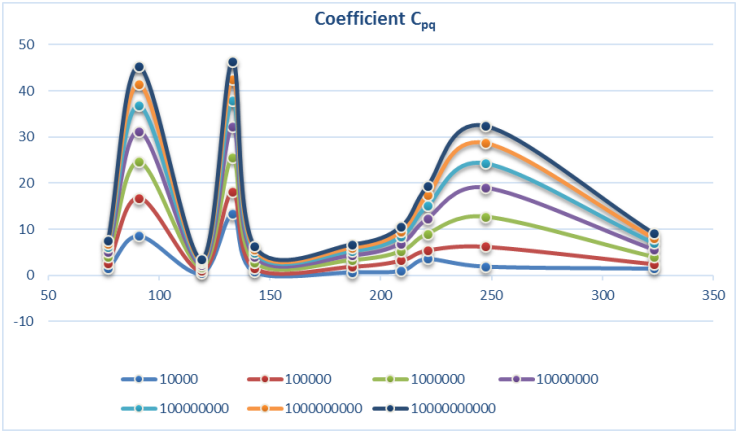

Figure 1. Estimation of the coefficient value (dependence on *pq*).

Figure 2 shows that the value of the coefficient increases monotonically, but does not exceed the theoretical value. So, we can make an assumption about approaching the theoretical value. However, due to the fact that there is no obvious slowdown in growth, it is difficult to conclude that there is better upper bound.

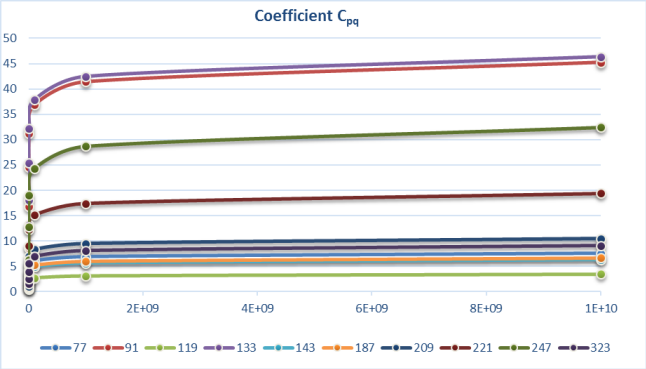

Figure 2. Estimation of the coefficient value (dependence on the boundary X).

Figure 3 shows that the value of the coefficient increases monotonically, but the growth rate gradually slows down. From this, we can make an assumption about the existence of an upper bound.

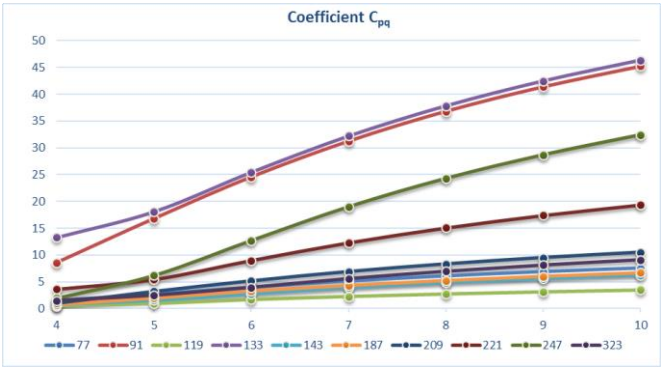

Figure 3. Estimation of the coefficient value (dependence on the logarithm of boundary X).

4.3. Function ratio

As we can see from the Figure 4 and Figure 5, the ratio between the functions very slowly approaches value 1, but does not reach. Also, it seems the limit of the sequence is not 1. It means that the resulting estimation is upper bound but not accurate and needs improvement.

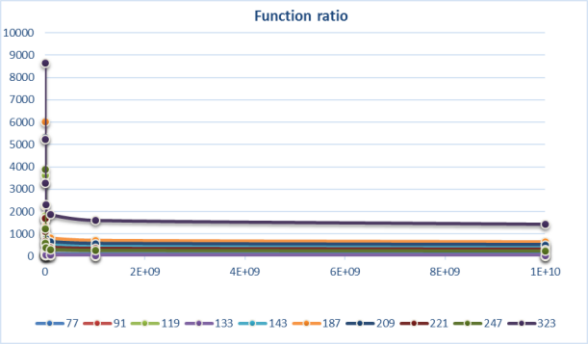

Figure 4. Function ratio (dependence on the boundary *X*, theoretical/practical).

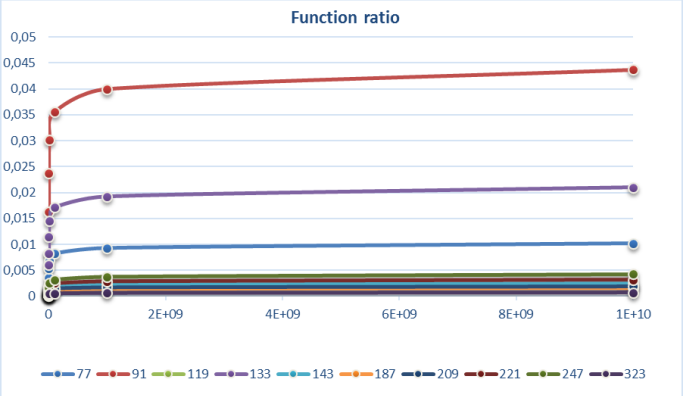

Figure 5. Function ratio (dependence on the boundary *X*, practical/theoretical).

4.4. Function comparison

We can see from the Fig. 6, which present function comparison for $pq = 77$, that the theoretical estimate is an upper bound for $Fr(n)$. We can notice that distance between the practical values and theoretical ones doesn't change much, so it can be assumed that the dependency type for the upper bound was found correctly. However, the theoretical function quite far from the practical one, which indicates an inaccurate finding of the coefficient C_{pq}.

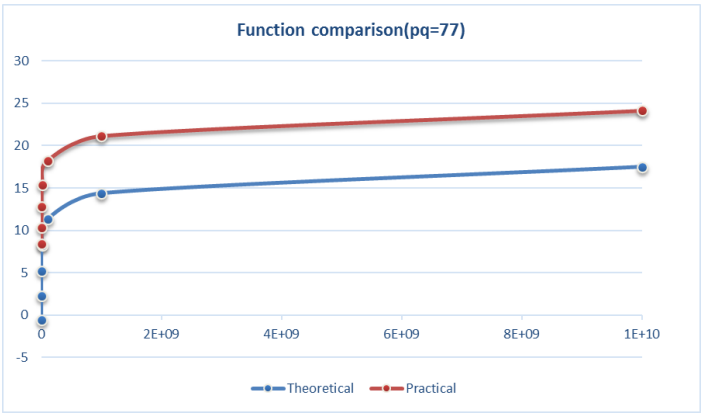

Figure 6. Comparison of theoretical and practical evaluation (dependence on the boundary *X*, *pq=77*).

4.5. Convergence to upper bound

We can see from the Fig. 7, that ratio very quickly reach value 1. It means that the resulting function is very accurate approximation of practical value. Also, we can see that after $\lg(X) = 7$ ratio become less than 1. It means the resulting function is upper bound for practical value.

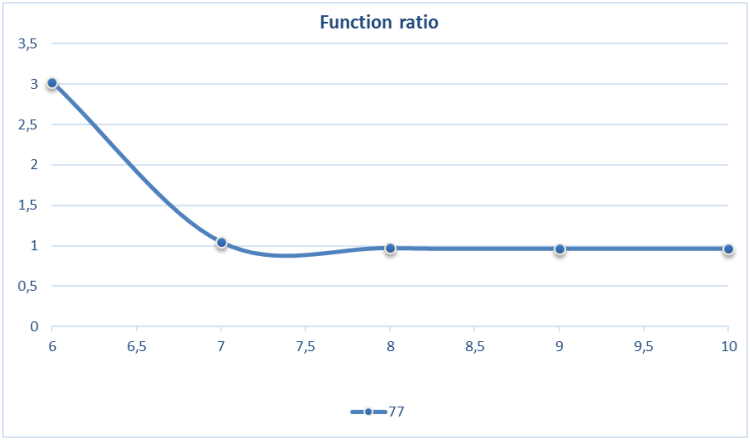

Figure 7. Function ratio (dependence on the logarithm of boundary *X*, practical/theoretical).

4.4. General conclusions

As we can see from comparison of theoretical and practical values, upper bound from (29) doesn't approximate $Avg(Fr(n))$ and needs improvement. However, dependency type for the upper bound as in (30) was found correctly. It means that the main improvement must be decrease of value C_{pq}.

But for subclass of tripleprime numbers with properties (31), (32), (33) and (34) we found very accurate approximation of $Avg'(Fr(n))$. One use of this estimation can be to improve the estimation for the whole class of tripleprime numbers.

5. CONCLUSIONS

In this article, we review current results for an upper bound for the average probability of error of the Miller-Rabin test. Also, we calculate new estimation for class tripleprimes numbers and made a comparison with practical results.

Conclusions were drawn about the correctness of the type of distribution of the theoretical estimation. However, it was found that the value of the coefficient C_{pq} is too high. All conclusions were accompanied by graphs for visual demonstration.

Also, we found very accurate approximation of the average probability of error of the Miller-Rabin test for some subclass of tripleprime numbers.

All our conclusions accompanied with graphs for more clarity.

Therefore, a further direction for investigation may be to attempt to decrease the value of the C_{pq} coefficient in order to obtain a more accurate upper bound.

ACKNOWLEDGEMENTS

The authors would like to thank everyone, just everyone!

REFERENCES

[1] Childs, N. Lindsay (2009) *A Concrete Introduction to Higher Algebra*, 3rd edition, Springer Publisher.
[2] Cormen, H. Thomas & Leiserson, E. Charles & Rivest, L. Ronald & Stein, Clifford (2001) *Introduction to Algorithms*, 2nd edition, MIT Press.
[3] Alford, W. R. & Granville, Andrew & Pomerance, Carl (1994) "There are Infinitely Many Carmichael Numbers", Princeton University *Annals of Mathematics*, Vol. 140, No. 3, pp703-722.
[4] Miller, G. (1976) "Riemann's hypothesis and tests for primality", Elsevier *Journal of Computer and System Sciences*, Vol. 13, pp300-317.
[5] Rabin, M. O. (1980) "Probabilistic algorithm for testing primality", Elsevier *Journal of Number Theory*, Vol. 12, No. 1, pp128–138.
[6] Nari, Kubra & Ozdemir, Enver & Ozkirisci, Neslihan Aysen (2019) *Strong pseudo primes to base 2*, arXiv Publisher, arXiv:1905.06447.
[7] Sorenson, P. Jonathan & Webster, Jonathan (2019) *Two Algorithms to Find Primes in Patterns*, arXiv Publisher, arXiv:1807.08777.
[8] Ribenboim, Paulo (1995) *The New Book of Prime Number Records*, 3rd edition, Springer Publisher.
[9] Zhang, Zhenxiang & Tang, Min (2003) "Finding strong pseudoprimes to several bases", American Mathematical Society *Mathematics of Computation*, Vol. 72, No. 244, pp2085-2097.
[10] Sorenson, Jonathan & Webster, Jonathan (2015) "Strong Pseudoprimes to Twelve Prime Bases", American Mathematical Society *Mathematics of Computation*, Vol. 86, No. 304, pp985-1003.
[11] Zhang, Zhenxiang (2007) "Two kinds of strong pseudoprimes up to 10^36", American Mathematical Society *Mathematics of Computation*, Vol. 76, No. 260, pp2095-2107.
[12] Ishmukhametov, S. T. & Mubarakov, B G. & Rubtsova, R G. (2020) "On the Number of Witnesses in the Miller–Rabin Primality Test", MDPI *Symmetry*, Vol. 12, No. 6, p890.
[13] Mubarakov, B G. (2020) "Efficient evaluation of the Miller-Rabin primality test of natural numbers", Kazan Mathematical Society *Proceedings of the N.I. Lobachevsky Mathematical Center*, Vol. 59, pp106-109.
[14] Mubarakov, B G. (2021) "On the Number of Primality Witnesses of Composite Integers", Springer *Russian Mathematics*, Vol. 65, No. 9, pp73–77.
[15] Hartley, R.V.L. (1928) "Transmission of information", IEEE The Bell System Technical Journal, Vol. 7, No. 3, pp535-563

AUTHORS

Zhumaniezov Alisher Assistant professor at Kazan Federal University. Graduated from Kazan Federal University and Czech State University in 2018. Has secondary job in the Laboratory of Medical Cybernetics and Machine Vision. Preferred areas: Cryptography, Machine vision.