Application of probabilistic analysis to the problem of finding the shortest route

Vasilyev V. V., Nikiforov K. A.

March 8, 2020

Abstract

In this article we review the application of the Pallottino algorithm to find the shortest path and investigate its complexity function using probabilistic analysis methods. The results are compared with the classical empirical estimation of the algorithm. A comparison of the Pallottino algorithm with its closest competitors on real geographical graphs is also given.

Keywords: shortest path, algorithm analysis, confidence complexity, geographical graph

E-mail addresses : vasar007@yandex.ru (Vasilyev V. V.), k.nikiforov@spbu.ru (Nikiforov K. A.)

1 Introduction

Practically significant analysis results of a specific algorithm is to obtain such information, that could give the prediction possibility of the resource cost required by the algorithm when solving problems of this troubled area.

Currently, the most widely used estimate is the complexity function on average, on the basis of which the time estimates of the program implementation of the algorithm can be predicted (in the statistical sense) with a sufficiently good accuracy.

However, the problem is that the average estimate, being statistical and point, does not provide any information about the behavior of the algorithm at specific inputs, which is important for both large-scale problems and real-time systems. The use of estimates in the worst case leads to a significant overestimation of the time forecast due to the low probability of inputs that provide maximum complexity at a fixed dimensionality of the problem. Thus, the problem of constructing a practically significant interval estimation of the algorithm complexity is of interest.

It is also worth noting the methods of machine learning, which has been actively improved in recent decades and are used in various fields, such as medicine, economics, computer vision, bioinformatics, information retrieval, etc. Their applications have been used to improve algorithms based on input data [11]. We assume that this method of research of algorithms deserves attention and will actively develop in the future, competing with classical approaches.

Another possible solution to the problem of improving the algorythm empirical analysis results accuracy is related to the consideration of the algorithm complexity at a fixed input length as a discrete bounded random variable having some unknown distribution. The approach consists in the construction of a confidence interval of labor intensity based on the approximation of the unknown discrete distribution of labor intensity values by a continuous distribution with limited variation, which is proposed to use the beta distribution. The resulting solution allows to specify a more real right boundary of the algorithm complexity at a fixed input length with a given confidence probability.

This method includes two stages — the stage of preliminary research, the purpose of which is to test the hypothesis about the law of distribution of the algorithm's labor intensity values as a discrete limited random variable [10], and the stage of the main study, which determines the values of the confidence labor intensity $f_\gamma(n)$ as a function of the input length of the algorithm [9].

As an algorithm for the study, we take one of the algorithms for finding the shortest paths in the graph — the Pallottino algorithm.

Finding the shortest path from one vertex to all the others is one of the basic network optimization problems. There are several common and well-known algorithms that solve this problem. One category consists of the Dijkstra algorithm [1] and its modifications, the other category contains the Bellman – Ford – Moore algorithm [2, 3, 4] and its modifications. There are also a number of other algorithms that are not

included in these categories, but they will not be reviewed by the authors.

Pallottino algorithm — the algorithm that finds the shortest distance from one of the vertices to all the others on graphs without loops, it is a modification of the Bellman – Ford – Moore algorithm. This algorithm also works for graphs with edges of negative weight under certain conditions. It is widely used to solve the problems of optimal distribution of cargo traffic on the transport network and the choice of the most profitable ways of its development.

In the literature there is a little confusion with the naming of this algorithm. Pape developed a suggestion of D'Esopo [**?**] and proposed improved algorithm [6]. At the same time Levit and Livshits developed their own version with the same idea [7]. Thus, we have D'Esopo – Pape – Levit algorithm which uses a deque to maintain the labeled vertices in the graph. Later Pallottino suggested using two queues instead of a deque to avoid the exponential complexity of the algorithm [8]. And the last one is the algorithm that we investigate using probabilistic analysis methods.

For the analysis of the Pallottino algorithm, we used a generator of random graphs and the graph of real geographical roads of St. Petersburg.

The organization of an experimental study of the algorithm complexity has led to the need to modify the source code of the implementation of the algorithm associated with the placement of the counter to determine the value of the number of performed basic operations at this input.

2 A preliminary study

The main purpose of the experimental study of the algorithm complexity is to obtain the values of the complexity function on average, depending on the length of the input. In this regard, the organization of an experimental study of the algorithm complexity leads to the necessity to solve the following problems:

- modification of the source code of the software algorithm implementation associated with the placement of the counter to determine the value of the number of performed basic operations at this input;
- the organization of the generation of the algorithm inputs to ensure representativeness of the sample, i. e. the generation of inputs corresponding to the features of the application of this algorithm in the studied software system;
- planning of the experimental study, that consists of determining the minimum required sample size at a fixed input length, determining the segment and the step of the length change.

As for the generation of the algorithm inputs, the main task is to ensure the representativeness of the sample, i. e. the generation of such inputs, which in probability correspond to the peculiarities of the application of this algorithm in the designed software system.

Let us examine the planning stage of the pilot study in greater detail [10]. The main objective of this stage is to determine the rational sample size at a fixed input length. In this paper, we assume that the complexity function of the Pallottino algorithm has beta distribution. The main stages of the preliminary study:

1. Fixing some value of the input length n from the real segment of lengths in the field of application of the Pallottino algorithm. In this case $n = 80$;
2. Determination of the required number of experiments m with software implementation to obtain a histogram of the relative frequencies of the labor intensity values, for example, by the method described in paper [10]. In this case $m = 52992$ (see section with the results);
3. Conducting experimental research and obtaining the values $f_i = f_A(D_i), i = \overline{1, m}$;
4. Theoretical functions of the algorithm complexity for best and worst cases, as functions of the input length. For the Pallottino algorithm, these functions have the form: $f_A^{\vee}(n) = n, f_A^{\wedge}(n) = (n-1)\frac{n+n(n-1)(n-2)}{6}$;
5. Select the number of semisegments for the frequency histogram of the labor intensity values. In this example, the histogram was based on 231 semisegments. This value for the number of semisegments was obtained by using the function of creating a histogram of relative frequencies from a standard package for data analysis in Microsoft Excel;
6. Normalization of the experimental values of intensity and build on the data obtained, the histogram of relative frequencies semisegments (figure 1);
7. Calculation of sample mean and sample variance from experimental data;
8. Formulation of the hypothesis and calculation of the parameters of the approximating distribution law. In this case, the hypothesis of beta distribution is put forward. The values of beta distribution parameters: $\alpha = 14.8044505, \beta = 1435255.779$;
9. Calculation of theoretical frequencies by density function (the results are shown in figure 1);
10. Calculation of the observed value of the Pearson criterion. In this case $\chi^2_{\text{obs}} = 0.13917862522660$;
11. To test hypotheses about the distribution law. If there is no reason to reject the null hypothesis, the transition to the main stage of the study is performed. Otherwise — make the choice of another distribution law and re-test the hypothesis. In this example, $\chi^2_{\text{cr}}(0.05, 228) = 265.30124341715, \chi^2_{\text{obs}} < \chi^2_{\text{cr}}(0.05, 228)$, and there is no reason to reject the null hypothesis. Also, the null hypothesis is confirmed by the standard function CHISQ.TEST from Microsoft Excel.

3 Basic research phase

1. Determination of the segment of the input length values corresponding to the peculiarities of the algorithm application in the developed software system. In this case, the Pallottino algorithm will be used for arrays from 80 to 2560;
2. The definition of the segment lengths of the entrance, which will be conducted in the pilot study. In this example, the segment is from 80 to 320;
3. Selection of the input length change step in the experimental study. In this case, the step value is 10;
4. Selection of the required number of experiments with the software implementation of the algorithm with a fixed input length, for example, by the method described in paper [10], to determine the sample mean and variance. In this case, $m = 52992$ (see the section with the results);
5. Calculation based on experimental data of sample mean and variance values for each n value. In this case, n changes from 80 to 320 in increments of 10;
6. Analysis of experimental data – construction of regression equation for sample variance. The results are shown in figure 2, and the regression equation is $s^2 = 0.1616n^{1.7125}$. In this case, $y = ax^b$ is the best in the sense of the maximum value of the coefficient of determination R^2;
7. Calculation based on the obtained results of the parameters of the approximating beta distribution as functions of the input length $\alpha(n), \beta(n)$. For the case under consideration, the graph of $\alpha(n)$ is shown in figure 3, and the graph of $\beta(n)$ is shown in figure 4. Also, the figures show the functions obtained by approximation;
8. Choosing the confidence value and calculating the values of the left γ quantile of the beta distribution: $x_\gamma(n) = B^{-1}(\gamma, \alpha(n), \beta(n))$. In this case, $\gamma = 0.95, \alpha \ll \beta$, the graph of $x_\gamma(n)$ is shown in figure 5;
9. The calculation of the function values of the trust complexity of the formula for the studied segment of the entrance lengths:

$$f_\gamma(n) = f^\vee(n) + x_\gamma(n)(f^\wedge(n) - f^\vee(n)).$$

Figure 6 shows the graph of the confidence and labor intensity in the worst case for the Pallottino algorithm on the segment [80, 2560]. It should be noted that the confidence complexity is obtained for the value of the confidence probability $\gamma = 0.95$, i. e. in 95% of cases with the probability observed in a single experiment the complexity of the algorithm will not exceed the value of the confidence complexity — for the example under consideration, these values are hundreds of thousands of times less than labor intensity in the worst case at small values ($n < 240$) of the investigated segment of the input lengths and millions of times less than labor intensity in the worst case at the other values ($n \geq 240$) of the investigated segment of the input lengths, as can be seen in figure 6.

4 Results

Для оценки необходимого числа экспериментов с программной реализацией алгоритма Паллоттино для фиксированной длины входа в соответствие с методикой, изложенной в [10], был выбран путь вычисления данного числа на основе нормального распределения с надёжностью $\gamma = 0.95$. Стоит отметить, что в работе [10] авторы показывают, что решение на основе бета-распределения даёт результат, который примерно на 40% меньше значения необходимого числа экспериментов решения на основе нормального распределения. Однако в данной работе был использован более простой вариант решения на основе нормального распределения.

Для определения необходимого числа экспериментов было проведено предварительное исследование алгоритма Паллоттино с объёмом выборки равным 200, вычислен коэффициент вариации V_f, и вычислено значение $m^*_{(1)}$, результаты приведены в таблице 1.

Предварительный объём выборки	200
Выборочное среднее	149.56
Выборочная дисперсия	317.77528
Выборочное отклонение	17.82625245
Коэффициент вариации	0.119191311
Рассчитанный объём выборки	54576

Таблица 1

Далее была извлечена выборка объёмом 54576, результаты её обработки приведены в таблице 2. Поскольку рассчитанный объём выборки оказался меньше, чем объём выборки текущего эксперимента, то $m = 53727$.

Предварительный объём выборки	54576
Выборочное среднее	148.20
Выборочная дисперсия	307.16872
Выборочное отклонение	17.52622944
Коэффициент вариации	0.118259959
Рассчитанный объём выборки	53727

Таблица 2

5 Conclusion

In this article we confirmed that the Pallottino algorithm for graphs with "geometric" origin, i.e. for graphs constructed on the basis of transport networks and real distances, works much faster than its asymptotic estimates.

Also we note that the estimation of the Pallottino algorithm complexity function in the worst case leads to a significant overestimation of the time forecast due to the low

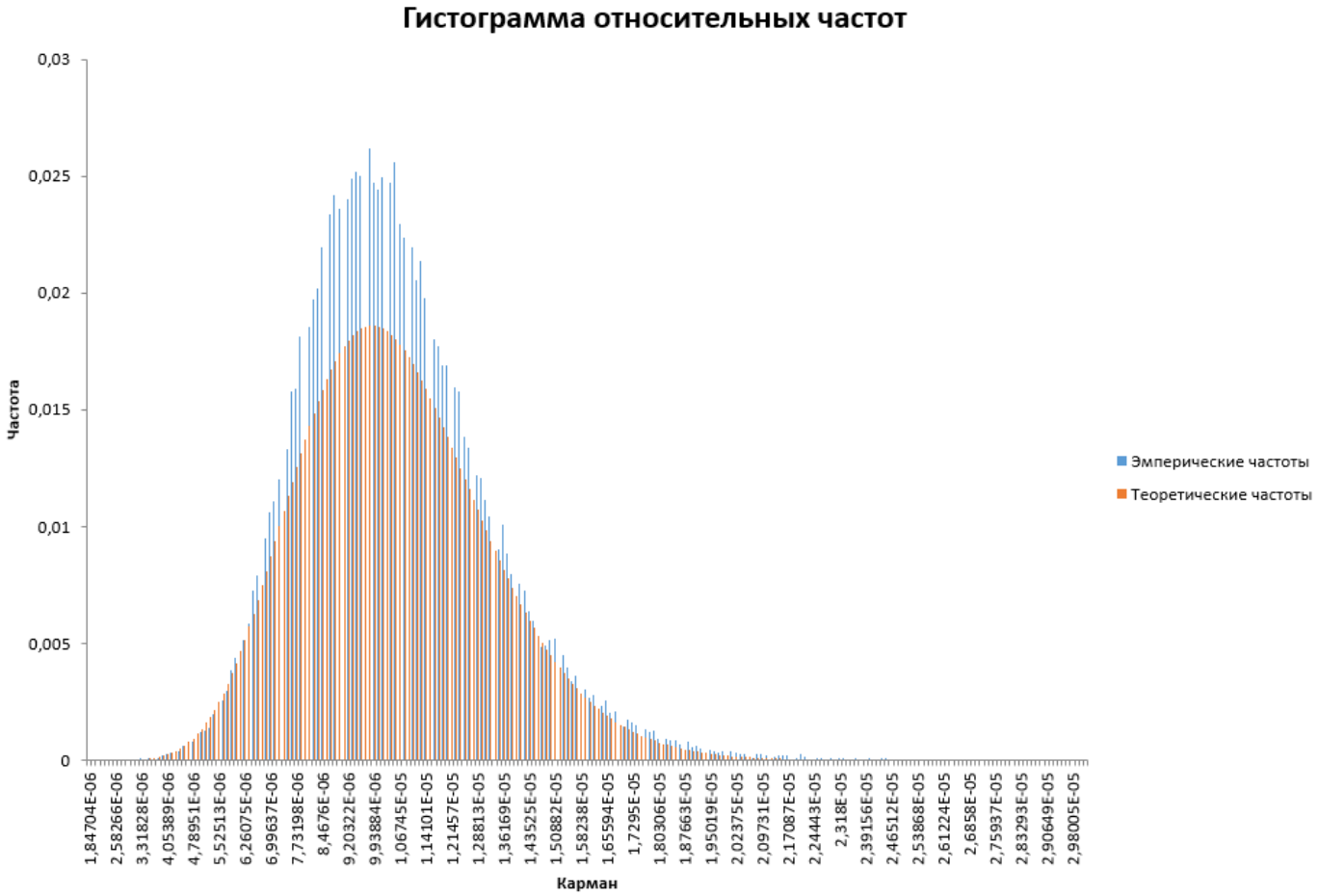

Figure 1: теоретические и эмпирические частоты для алгоритма Паллоттино при $n = 80$ с разбиением нормированного сегмента [0, 1] на 232 полусегмента

probability of inputs that provide maximum complexity function for the considered dimensionalities of the problem to find the shortest path in the graph.

The obtained results confirm the possibility of improving the reliability of forecasting the time efficiency of computer algorithms and a better solution to the problem of choosing rational algorithms based on a comparative analysis of the functions of confidential labor intensity instead of the traditional comparison of labor intensity in the average case.

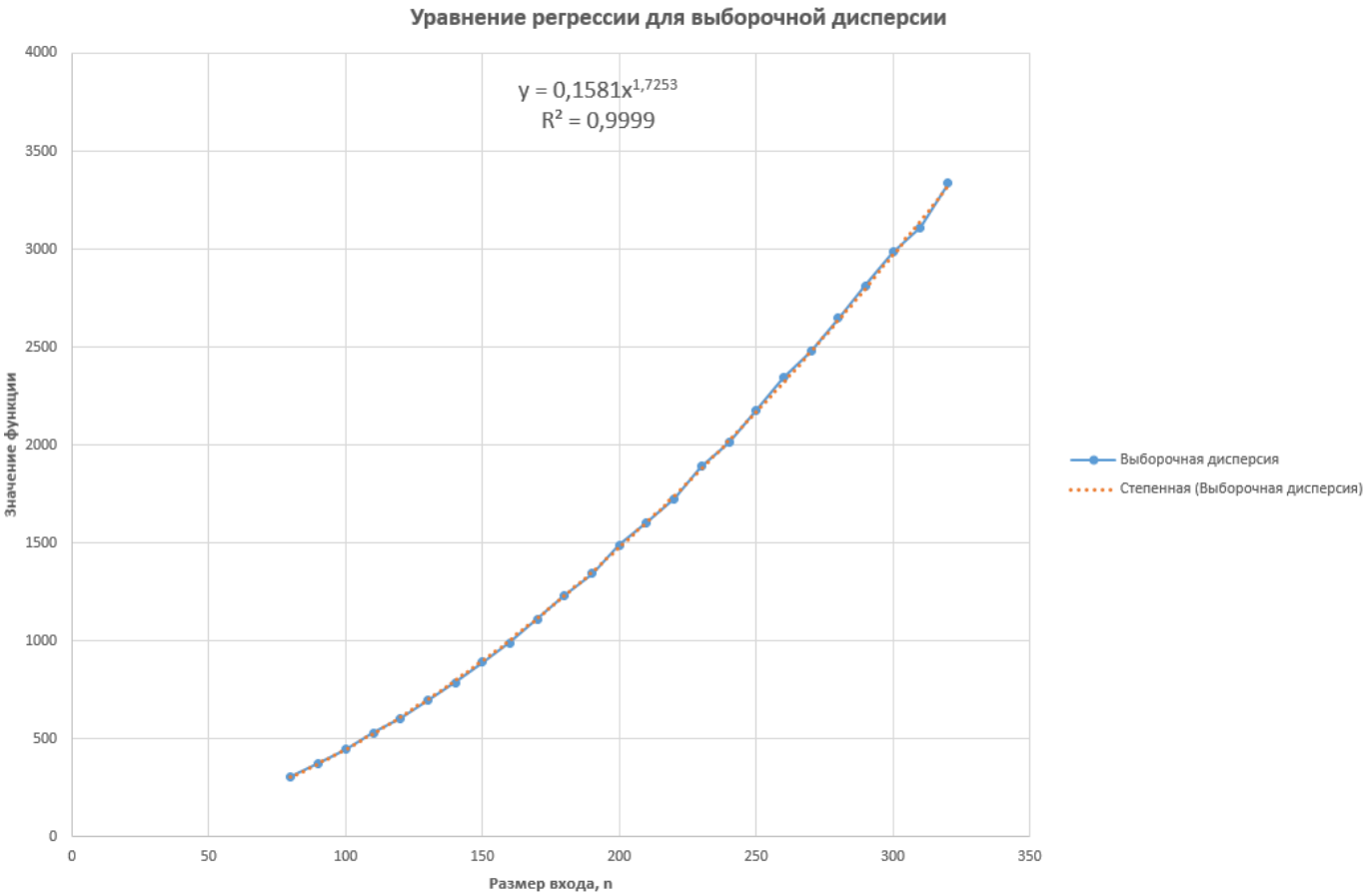

Figure 2: экспериментальные данные и уравнение регрессии для выборочной дисперсии значений трудоёмкости алгоритма Паллоттино

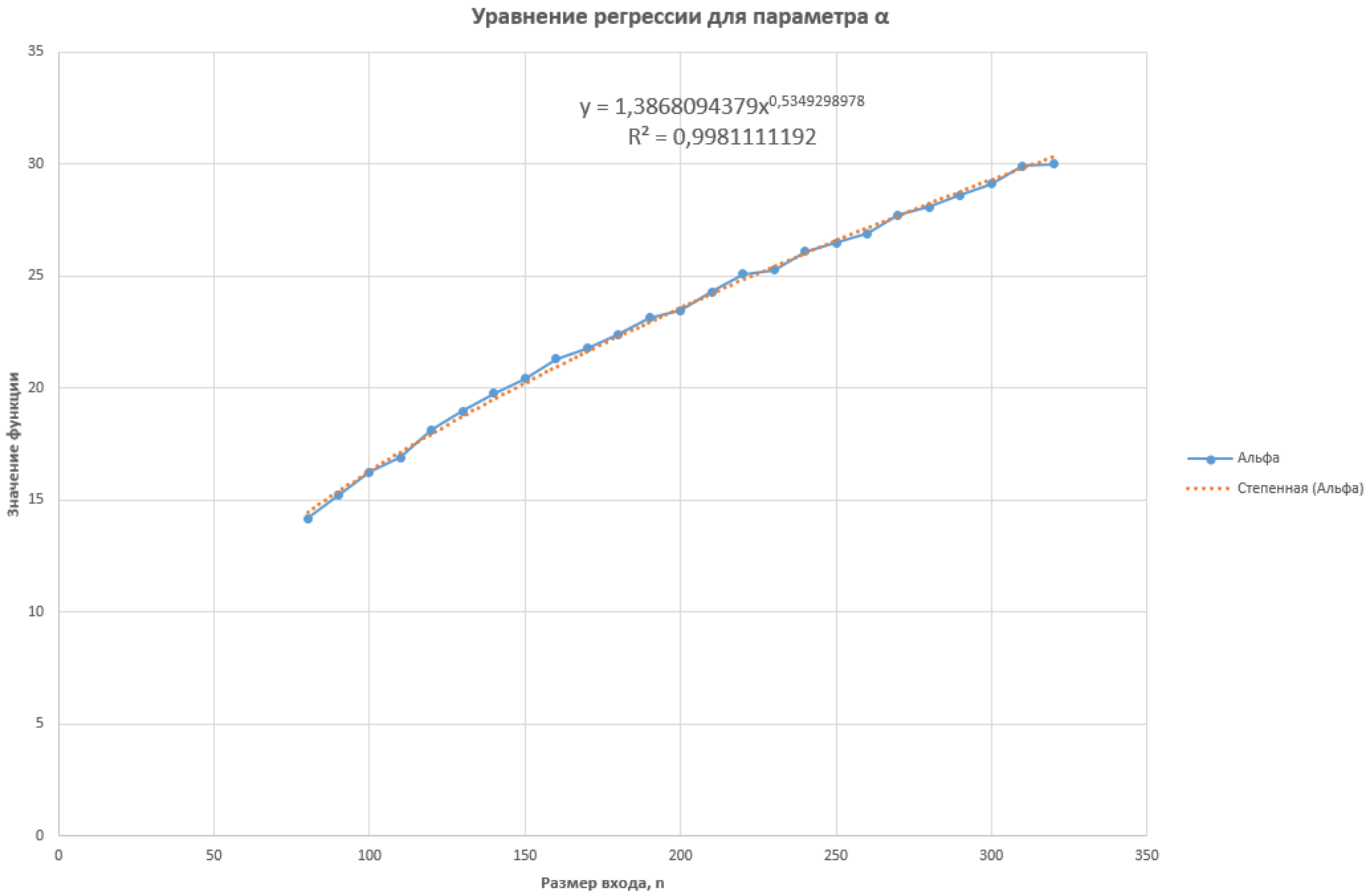

Figure 3: график функции $\alpha(n)$ — параметра α аппроксимирующего бета-распределения для Паллоттино

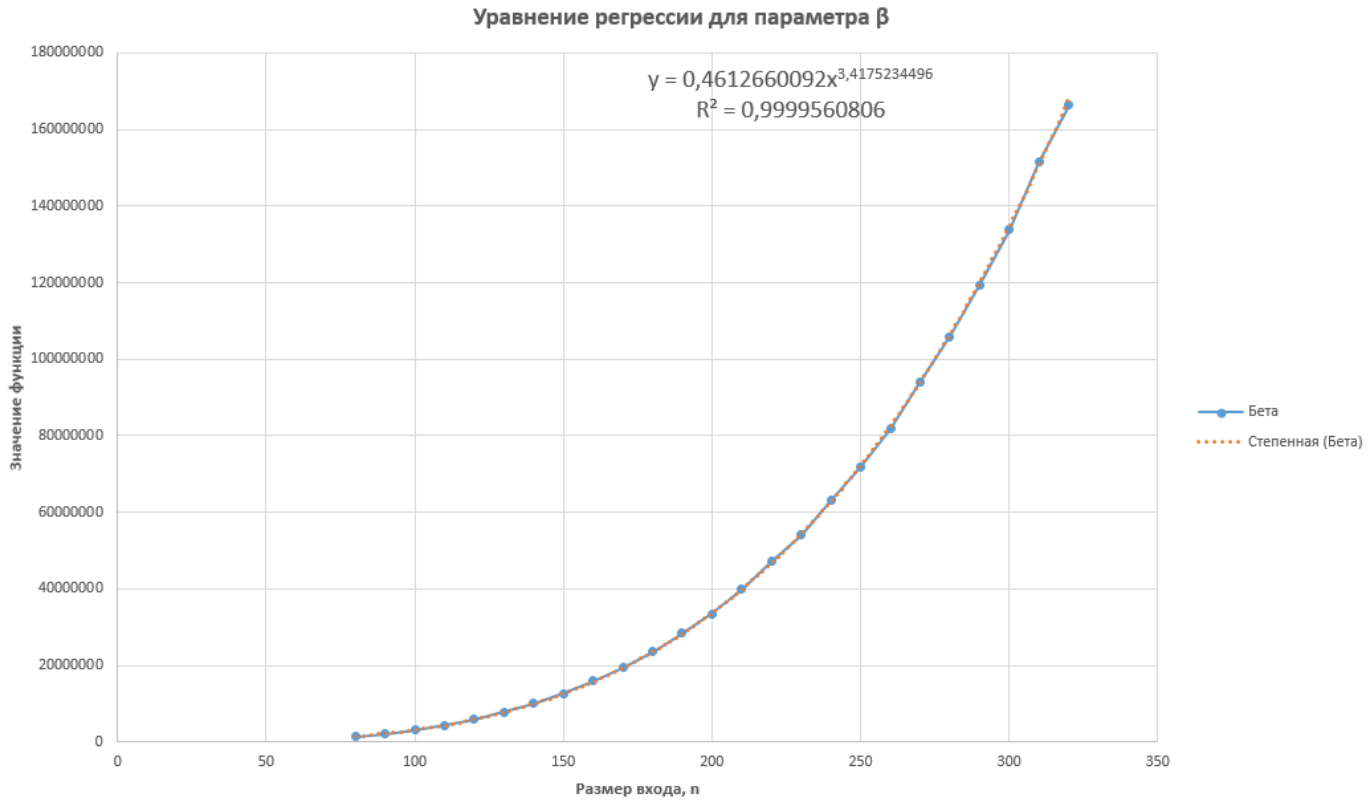

Figure 4: график функции $\beta(n)$ — параметра β аппроксимирующего бета-распределения для Паллоттино

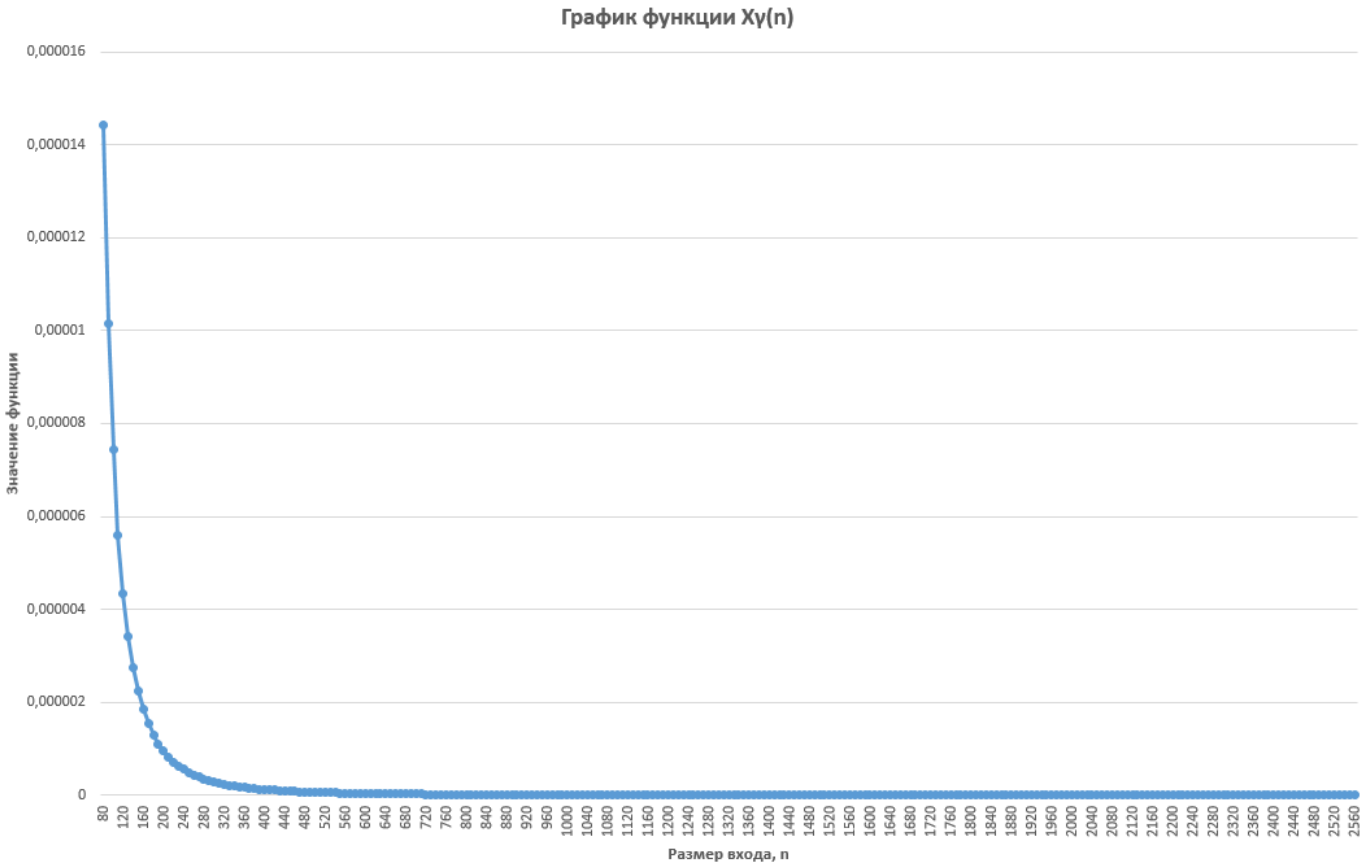

Figure 5: График зависимости левого γ-квантиля бета-распределения $x_\gamma(n)$ от длины входа для алгоритма Паллоттино

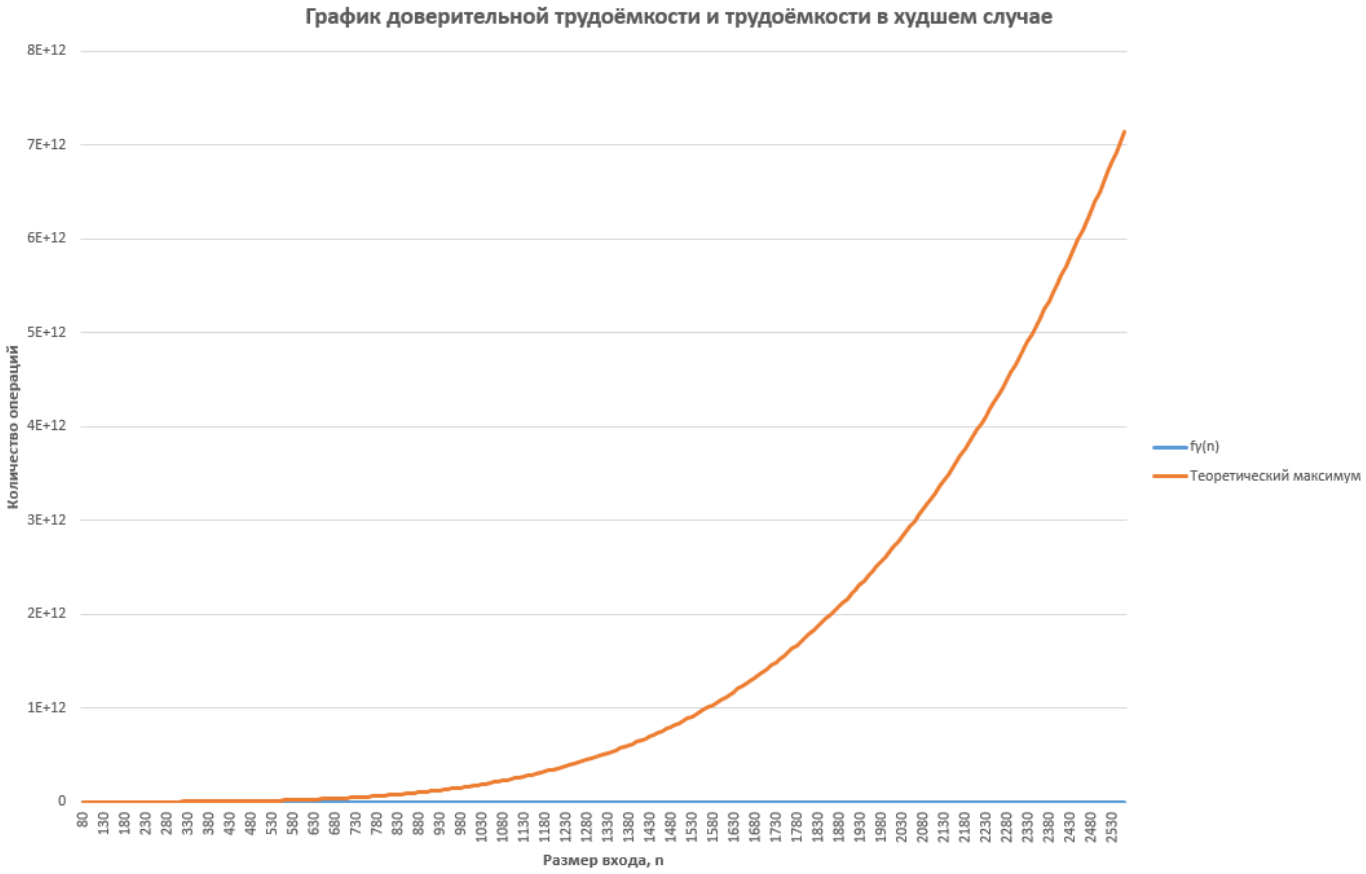

Figure 6: график доверительной трудоёмкости и трудоёмкости в худшем случае для алгоритма Паллоттино

6 References

1. E.W. Dijkstra, A note on two problems in connexion with graphs, Numer. Math. 1 (1959), 269–271.
2. R. Bellman, On a routing problem, Quart. Appl. Math. 16 (1958), 87–90.
3. L.R. Ford and D.R. Fulkerson, Flows in Networks, Princeton University Press, Princeton, NJ, 1962.
4. E.F. Moore, The shortest path through a maze, volume 3523 of Bell Telephone System. Technical publications. monograph, 1959.
5. M. Pollack and W. Wiebenson, Solutions of the shortest-route problem – A review, Oper. Res. 8 (1960), 224–230.
6. U. Pape, Implementation and efficiency of Moore-algorithms for the shortest route problem, Math. Program. 7 (1974), 212–222.
7. B.J. Levit and B. Livshits, Neleneinye setevye transportnye zadachi, Transport, Moscow, 1972.
8. S. Pallottino, Shortest-path methods: Complexity, interrelations and new propositions, Networks 14 (1984), 257–267.
9. Петрушин В. Н., Ульянов М. В, Кривенцов А. С., Доверительная трудоемкость — новая оценка качества алгоритмов // Информационные технологии и вычислительны

системы, 2009. №2. С. 23–37.

10. Петрушин В. Н., Ульянов М. В. Планирование экспериментального исследования трудоемкости алгоритмов на основе бета"=распределения // Информационные технологии и вычислительные системы, 2008. №2. С. 81–91.

11. P. Berube, J.N. Amaral, Combined profiling: A methodology to capture varied program behavior across multiple inputs, (2012) ISPASS 2012 - IEEE International Symposium on Performance Analysis of Systems and Software, art. no. 6189227, pp. 210–220.

A Appendix

1. Репозиторий проекта по анализу алгоритма Паллоттино на языке программирования C++
(https://github.com/Vasar007/algorithm_analysis).