May, 1996, IFTUWr 908/96

Unexpected Regularities in the Distribution of Prime Numbers

Marek Wolf

Institute of Theoretical Physics, University of Wrocław
Pl.Maxa Borna 9, PL-50-204 Wrocław, Poland, e-mail: mwolf@ift.uni.wroc.pl

Abstract

The computer results of the investigation of the number $h_N(d)$ of gaps of the length d between two consecutive primes smaller than N are presented. The computer search was done up to $N = 2^{44} \approx 1.76 \times 10^{13}$. The oscillations of $h_N(d)$ are found; the most eminent have the period 6. A few attempts to fit the obtained data by an analytical formula are given. As the applications two formulae are obtained: for the largest gap between consecutive primes below a given bound and the formula for "champions" — the most often occurring pairs of primes. Obtained data supports the conjecture that the number of Twins and primes separated by a gap of the length 4 ("Cousins") is almost the same and it determines a fractal structure on the set of primes. The possible relevance of the obtained results for the quantum chaos is pointed out.

Key words: *Prime numbers, gaps between primes, quantum chaos*

"...upon looking at prime numbers one has the feeling of being in the presence of one of the inexplicable secrets of creation."
D.Zagier, in [6]

1 Introduction

Recently there appeared a few papers where it was shown how some mathematical facts about primes can be applied to the study of quantum chaos [1]. Among others, the Hardy–Littlewood conjecture [2] on the gaps between (not necessary consecutive) primes inspired a few works on the correlations between periodic orbits in some chaotic systems, see [4]. In this paper we are going to look for the statistical properties of the distribution of gaps between *consecutive* primes p_n, p_{n+1} (statistics of "nearest–neighbour spacings").

The problem of the appearance of gaps between consecutive primes has a long history, see e.g. [3], [5], [6], [8], [9], [10]. These authors were mainly interested in the first appearance of a given gap of length d and here we are interested in the statistical properties of the distribution of gaps.

The main quantity studied here is defined as follows:

$$h_N(d) = \text{number of pairs } \; p_n, p_{n+1} < N \text{ with } d = p_{n+1} - p_n. \tag{1}$$

I have made the computer search for prime numbers up to $N = 2^{44} \approx 1.76 \times 10^{13}$ and counted the number of gaps between consecutive primes [11]. The computer program was implemented on the DEC Alpha 300X/175MHz workstation in DECFortran which allows some assembly level instructions and it took 132 days of CPU time to reach 2^{44}. The obtained data allow to make conjectures on the form of the dependence of $h_N(d)$. The formulae presented in this paper are heuristic, but nevertheless are in perfect agreement with conjectures of other authors [2], [8], [9] and/or are very well confirmed by numerical calculations. There are no theorems nor proofs — only the computer results, conjectures and interpretation of the observed regularities given by a physicist.

2 Exponential decreasing of $h_N(d)$ and oscillations

During the computer search the data representing the function $h_N(d)$ were stored at values of N forming the geometrical progression with the ratio 4, i.e. at $N = 2^{20}, 2^{22}, \ldots, 2^{42}, 2^{44}$. The results are plotted in the Fig.1. A lot of regularities can be seen on this Figure: the "peaks" and local minima appear in the same regular manner for each N and the pattern of circles *does not* depend on N — only the slope decreases.

Because the points lie around the straight lines on the semi-logarithmic scale, we can infer from the Fig.1 the conclusion that $h_N(d)$ decreases exponentially with d:

$$h_N(d) \sim B_1(N) e^{-A_1(N)d}. \tag{2}$$

To determine the slope $A_1(N)$ and the intercept $B_1(N)$ let me remark, that there are two selfconsistency conditions that the functions $A_1(N)$ and $B_1(N)$ have to obey. First of all, the number of all gaps is by 1 smaller than the number of primes:

$$\sum_d h_N(d) \;=\; \pi(N) - 1, \tag{3}$$

where $\pi(N)$ denotes the number of primes smaller than N. The second selfonsistency condition comes from the observation, that the sum of distances between consecutive primes $p_n \leq N$ is equal to the largest prime $\leq N$. So, for large N we can write:

$$\sum_d h_N(d)\, d \;\approx\; N. \tag{4}$$

For the Gauss estimation $\pi(N) \sim N/\ln(N)$ from these identities, after summing up geometrical series, it follows, that the functions $A_1(N)$ and $B_1(N)$ have the form:

$$A_1(N) \sim \frac{1}{\ln(N)}, \qquad B_1(N) \sim \frac{cN}{\ln^2(N)}, \qquad c = 2. \tag{5}$$

However the points in the Fig.1 very clearly oscillate around the lines being the best least-square fits. In fact, the period of this oscillation is exactly 6 plus less significant larger periods, see Fig.2. I claim that these oscillations are described by the following modification of the formula (2):

$$h_N(d) \sim B_2(N) \prod_{p|d, p>2} \frac{p-1}{p-2} e^{-A_2(N)d}, \tag{6}$$

where $p \mid d$ denotes such primes p which divide d. The product

$$\prod_{p|d, p>2} \frac{p-1}{p-2} \tag{7}$$

appearing in front of the exponent in (6) I have borrowed from the n-tuple conjecture of Hardy and Littlewood [2]. In the Fig.2 the power spectrum obtained via the discrete Fourier Transform of $h_N(d)/B_1(N)e^{-A_1(N)d}$ as well as of the product (7) is shown. Because the two curves in the Fig.2 have the peaks exactly at the same periods, all regularities and oscillations of different periods seen in the Fig.1 are caused by the product (7). The points obtained from formula (6) reproduce quite well data obtained from the computer search, see [11]. In the physical language the formula (2) can be regarded as a zero order (in the periods of oscillations) approximation to the formula (6). The “first order” approximations, taking into consideration only the period 6, are discussed in [11]. The problem is that for (6) I am not able to write down in the closed form the selfonsistency conditions. But the product (7) behaves very erratically and takes values of the order 1; e.g. for the $N < 2^{44}$ it takes values between 1 and 3.2 (3.2 appears only twice: for $d = 210$ and $d = 420$). So I suggest that the functions $A_2(N)$ and $B_2(N)$ are also given by eq.(5) but with different constant c. The value of c for (6) can be obtained by the following reasoning: To calculate the sums in (3) and (4) for $h_N(d)$ given by (6) the product (7) can be replaced by its mean value s defined as:

$$s = \lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} \prod_{p|2k, p>2} \frac{p-1}{p-2} \tag{8}$$

Then for the slope and intercept I get: $A_1(N) \sim 1/\ln(N)$, $B_1(N) \sim 2N/s\ln^2(N)$. This average value s of the product (7) can be determined from the consistency with the Hardy-Littlewood conjecture [2] for the number of Twins:

$$h_N(2) \sim c_2 \frac{N}{\ln^2 N}, \tag{9}$$

where the constant

$$c_2 \equiv 2\prod_{p>2}\left(1 - \frac{1}{(p-1)^2}\right) = 1.32032\ldots. \tag{10}$$

Assuming that N is so large that [1] $\exp(-2A(N)) \approx 1$, for the number of Twins from (6) I get $h_N(2) \sim 2N/s\ln^2(N)$ and comparison with (9) gives:

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} \prod_{p|2k, p>2} \frac{p-1}{p-2} = \frac{1}{\prod_{p>2}(1 - \frac{1}{(p-1)^2})} \quad (= \ 1.51478\ldots). \tag{11}$$

This equality really holds, as can be proved rigorously [2] by an application of the Ikehara–Delange theorem [7], so it supports the validity of our conjectures. These arguments lead me to the suggestion, that $A_2(N) \sim 1/\ln(N)$, $B_2(N) \sim c_2N/\ln^2(N)$ and finally the main conjecture of this paper reads:

$$h_N(d) \sim \frac{c_2 N}{\ln^2(N)} \prod_{p|d, p>2} \frac{p-1}{p-2} e^{-d/\ln(N)}. \tag{12}$$

3 Two Applications

As the application of the formula (2) we can obtain the length $G(N)$ of the largest gap between consecutive primes below a given bound N. Simply, the largest gap G appears only once, so it is equal to the value at which $h_N(d)$ crosses the d-axis on the Fig.1: $h_N(G(N)) = 1$; it gives:

$$G(N) \sim \ln(N)(\ln(N) - 2\ln\ln(N) + c_0), \tag{13}$$

where $c_0 = \ln(2) = 0.6931\ldots$ for (2) and $c_0 = -\ln(\frac{1}{2}c_2) = 0.4153\ldots$ for (6).

The above formula agrees well with the actual computer data, see [11] and for large N passes into the Cramer [8] conjecture (see also [9]):

$$G(N) \sim \ln^2(N). \tag{14}$$

[1] The approximation $\exp(-dA(N)) \approx 1$ for large N is the best one for $d = 2$ and worsens with increasing d.

[2] W.Narkiewicz: personal communication

The examination of the eqs.(13) and (14) with real data is shown in the Fig.3, which should be compared with the figure on the page 12 in [6].

As an another application of the formula (6) let us consider the most often occurring distances between primes, called "high-jumpers" or "champions", [12]. The straightforward checking shows, that up to $N = 389$ the largest number of gaps is for $d = 2$, while the most often one becomes the gap $d = 6$ — the circle corresponding to $d = 6$ is the highest one on the plots for each N in the Fig.1. In the Fig.1 the local spikes appear at multiplicities of $30 = 2{\cdot}3{\cdot}5$. For larger values of N even higher spikes at $d = 210 = 2{\cdot}3{\cdot}5{\cdot}7$ can be seen. As N increases the slopes $A(N)$ decrease and at some value $N^{(3)}$ the peak at $d = 30$ will be greater than that at $d = 6$. At much larger $N^{(4)}$ the spike at $d = 210$ will take over $d = 30$. From this observation and eq.(12)I can obtain the general estimation for the values of $N^{(n)}$ at which the consecutive products $D^{(n)} = 2 \cdot 3 \cdot \ldots \cdot p_n$ become the champions:

$$\ln(N^{(n)}) = \frac{2 \cdot 3 \cdot \ldots \cdot p_{n-1}(p_n - 1)}{\ln((p_n - 1)/(p_n - 2))}, \qquad n > 1. \tag{15}$$

Let me notice that the dependence on $B_2(N)$ drops out and the form of (15) *does not* depend on the particular value of the constant c_2. The formula (15) can be justified in more rigorous manner, see [12]. The values of $N^{(n)}$ obtained from (15) are shown in the Table 1. These numbers grow pretty fast and can compete with such large numbers as those of Skewes, Graham, Paris-Harrington or Friedman, see [13]. Obtained from (15) for $D^{(2)} = 6$ the value of $N^{(2)} = 321$ quite well agrees with the actual value 389 (twins $D^{(1)} = 2$ are champions simply from the begining).

$D^{(n)}$	$N^{(n)}$
6	3.21×10^{2}
30	1.70×10^{36}
210	5.81×10^{428}
2310	1.48×10^{8656}
30030	1.30×10^{138357}
510510	8.02×10^{3233259}
9699690	8.50×10^{69820169}
223092870	$5.14\times10^{1992163572}$
6469693230	$3.56\times10^{74595540317}$
200560490130	$2.10\times10^{2486392448589}$
$\vdots$	$\vdots$

Table 1.

N	$h_N(2)$	$h_N(4)$	$h_N(2)/h_N(4)$
2^{20}	8535	8500	1.0041176
2^{22}	27995	27764	1.0083201
2^{24}	92246	91995	1.0027284
2^{26}	309561	309293	1.0008665
2^{28}	1056281	1057146	0.9991818
2^{30}	3650557	3650515	1.0000115
2^{32}	12739574	12740283	0.9999443
2^{34}	44849427	44842399	1.0001567
2^{36}	159082253	159089620	0.9999537
2^{38}	568237005	568225073	1.0000210
2^{40}	2042054332	2042077653	0.9999886
2^{42}	7378928530	7378989766	0.9999917
2^{44}	26795709320	26795628686	1.0000030

Table 2.

4 Twins and Cousins

One of the most famous open problems in the number theory (see e.g [6]) is the Twin Primes conjecture, which states that there are infinitely many primes p with $p+2$ also a prime. The numbers of Twins are represented by a first circle $d=2$ on each plot (i.e. for every N) in the Fig.1. However, the next circles, corresponding to $d=4$, are located on the same attitude as Twins, see also Table 2. It means, that numbers of pairs of primes separated by the distance $d=2$ and $d=4$ are almost the same:

$$h_N(2) \approx h_N(4). \tag{16}$$

By analogy with Twins a pair of primes separated by a gap $d=4$ I will name "Cousins". As it is seen from the Table 2 the ratio $h_N(2)/h_N(4)$ is sometimes larger than 1, and for other N is smaller than 1. It means that there is a set of such N that the numbers of Twins and Cousins are equal. Because $h_N(d)$ can change value only for N being prime, it is reasonably to look for such primes $p^{(z)}$ that the difference $h_{p^{(z)}}(2)-h_{p^{(z)}}(4)$ is zero. I have checked [14] by direct computer search that up to $N=2^{43}$ there are 2823290 primes $p^{(z)}$, at which the numbers of Twins and Cousins are exactly the same. The distribution $\mathcal{D}_N(\Delta z)$ of spacings Δz between consecutive $p^{(z)}<N$ is shown in the Fig.4. There are 2790362 spacings shown on this figure — only 32928 distances between consecutive $p^{(z)}$ were larger than 10^4. In fact these larger distances have very scattered values (the largest gap between two clusters of the same number of Twins and Cousins was 314267840234) and they appeared only once — it resembles the intermittent behaviour in some dynamical systems. There is a cross-over from the exponential dependence of $\mathcal{D}_N(\Delta z)$ to the power-like decrease. It means that the set of $p^{(z)}$ is formed by "clusters" or "islands" separated by distances obeying the power law, what in turn is characteristic for fractal sets [15]. In other words, these clusters of $p^{(z)}$ are organized in a hierarchical, selfsimilar set. Because inside the islands (clusters) $h_{p^{(z)}}(2)$ and $h_{p^{(z)}}(4)$ do not change their values, spacings Δz between $p^{(z)}$ inside clusters follow the usual rule (6) and it is the explanation for the left part of the plot in the Fig.4. However, in the region $\Delta z \in (z_0, 360)$, where z_0 is roughly 120, the two behaviors overlap and it causes big 'jumps' of the values of $\mathcal{D}_N(\Delta z)$ at $\Delta z = 6n$, because all Twins are of the form $6n \pm 1$, (n are not necessary consecutive integers). But it is a little bit difficult to split the value of $\mathcal{D}_N(\Delta z)$ into the part arising from the distribution of gaps between consecutive primes and distances between "clusters" of primes at which $h_{p^{(z)}}(2) = h_{p^{(z)}}(4)$ holds. The exponent γ in the power–like part $\mathcal{D}_N(\Delta z) \sim (\Delta z)^{-\gamma}$ has the value of $\gamma \approx 1.46$ and it seems that this "critical exponent" *does not* depend on N, see Fig.4.

The possible existence of some hidden correlations among prime numbers and its influence for the quantum chaos was anticipated by M.C. Gutzwiller a few years ago, see popular exposition in [16]. Relevance of the regularities in the behaviour of $h_N(d)$ as well as the the intermittent and fractal behaviour of $p^{(z)}$ for quantum chaos is now under study. In particular, the next–to–leading terms to the two-point correlation function for the critical zeros of of the $\zeta(s)$ function should be possible to obtain. Maybe the formula (6) will provide also an explanation for the geometric regularities found "by eyes" by S.Ulam in the distribution of primes, see [17].

Acknowledgments. I thank prof. A.Granville and prof. A.M. Odlyzko for e-mail feedback and Prof. W.Narkiewicz for discussions. The author receives support from the KBN research grant No 2 P302 057 07.

References

[1] M. Berry, "Some quantum-to-classical asymptotics", in *Chaos and Quantum Physics*, edts. A.Voros *et al* (Elsevier, Amsterdam, 1991); R.Aurich et al, "Universal Signatures of Quantum Chaos", *Phys.Rev.Lett.* **73** (1994), p.1356; "Quantum Chaos", CXIX 'Enrico Fermi' International School of Physics, edts. G.Casati et al (North Holland, Amsterdam, 1993)

[2] G.H.Hardy and J.E. Littlewood, "Some problems of 'Partitio Numerorum' III", *Acta Mathematica* **44** (1922), p.1-70

[3] Polya, Amer.Math.Monthly, **66** (1959), p.375-384

[4] N.Argaman et al, "Correlations in the Actions of Periodic Orbits Derived from Quantum Chaos", *Phys.Rev.Lett.* **71** (1993), p.4326-4329 ; E.Bogomolny and P.Lebœuf, "Statistical Properties of the Zeros of Zeta Functions — Beyond the Riemann case", *Nonlinearity* **7** (1994), p.1155-67

[5] R.P.Brent,"The Distribution of Small Gaps Between Succesive Primes", *Math.Comp.* **28** (1974), p.315-324; R.P.Brent, "The First Occurrence of Certain Large Prime Gaps", *Math.Comp.***35** (1980), p.1435-1436

[6] D.Zagier, "The first 50 Million Prime Numbers", *Math.Intellig.* **0** (1977), p.7-19

[7] W. and F. Ellison *Prime Numbers* (John Wiley and Sons, New York, London, 1985)

[8] H.Cramer, "On the order of magnitude of difference between consecutive prime numbers", *Acta Arith.* **2** (1937), p.23-46

[9] D.Shanks, "On Maximal Gaps between Succesive Primes", *Math.Comp.* **18** (1964), p.464

[10] H.Maier, C.Pomerance, "Unusually large gaps between primes", *Trans. Amer. Math.Soc.* **322** (1990), p.201-237

[11] M.Wolf, preprint IFTUWr 894/95, submitted to Journ. Reine u. Angew. Math., available via anonymous ftp from orchid.ift.uni.wroc.pl (156.17.88.118), directory pub/primes.

[12] A.M. Odlyzko, M.Rubinstein, M.Wolf, to appear

[13] G.Kolata, "Does Göedel's Theorem Matter to Mathematics?", *Science* **218** (1982), p. 779

[14] M.Wolf, to be published; the file containing zeros of the difference $h_N(2) - h_N(4)$ for N being prime numbers up to 2^{43} is over 37MB long.

[15] B.B.Mandelbrot, *The Fractal Geometry of Nature*, (W.H.Freeman and Co., San Francisco, 1982)

[16] M.C.Gutzwiller, "Quantum Chaos", *Scientific American*, January 1992; see also J.Keating, "Physics and the Queen of Mathematics", *Physics World*, April 1990, p. 46

[17] A.K.Dewdney, "How to pan for primes in numerical gravel", *Scientific American.* July 1988, p.90-93

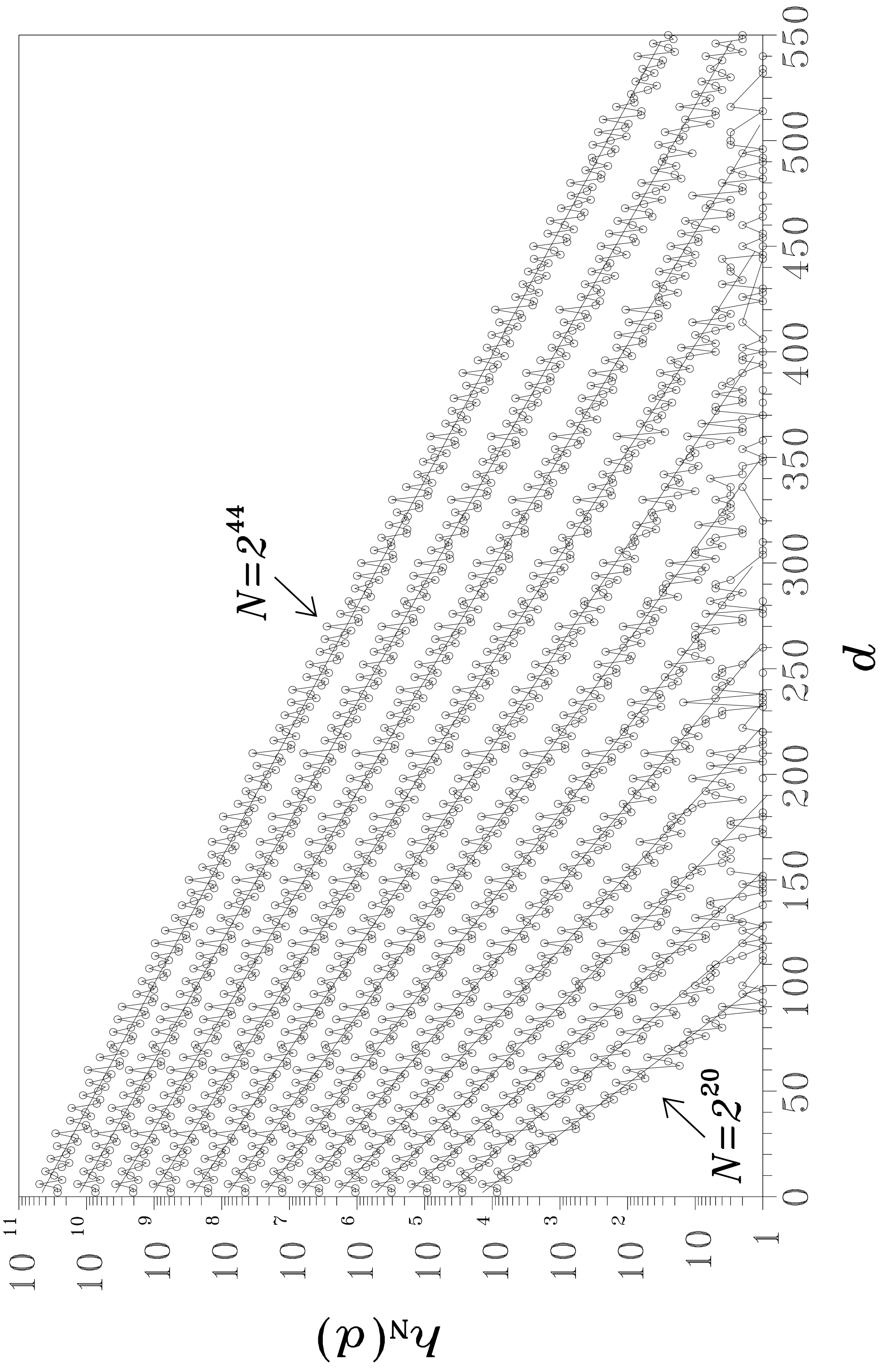

Fig.1 The plot showing the dependence of the histogram $h_N(d)$ on d at $N = 2^{20}, 2^{22}, \ldots, 2^{44}$. There is a logarithmical scale on the y-axis , while on the x-axis there is a linear scale. The values of $h_N(d)$ obtained from the computer search are represented by small circles. The straight lines are the best fits obtained by means of the least-square method. The points oscillate around the straight lines with period 6. Let us mention "local" spikes at $d = 30(= 2 \cdot 3 \cdot 5), 60, \ldots$. Especially well profound are spikes at $d = 210 = 2 \cdot 3 \cdot 5 \cdot 7$, and for $N = 2^{40}$, $N = 2^{42}$ and $N = 2^{44}$ also at its multiplicity $d = 420$ (second harmonic). More insights into the structure of the distribution of circles can be gained when looking at sliding angles.

FIG.2

Fig.2 The power spectrum as a function of the periods T for the oscillating part of $h_{N=2^{44}}(d)$ (after removing the exponential decrease and the prefactor $B(N)$) and for the product (7). The dashed line shows the power spectrum of the product (7) calculated for even arguments $d = 2k$ taken up to $d = 512$, to coincide with data for $h_N(d)$. The peaks at $d = 6$ were cut at 1000 to make visible much less maxima at $d = 30$ — the power spectrum for $h_N(d)$ took in fact the value over 2500 and over 6000 for the product. The horizontal axis starts at $T = 4$, because the Nyquist frequency for this case is 1/4. Although at $T = 30$ the peak is rather small, it adds up to the periods 5, 6, 10 and 15 dividing 30, what causes the emergence of spikes at $d = 30k$ seen in Fig.1.

FIG.3

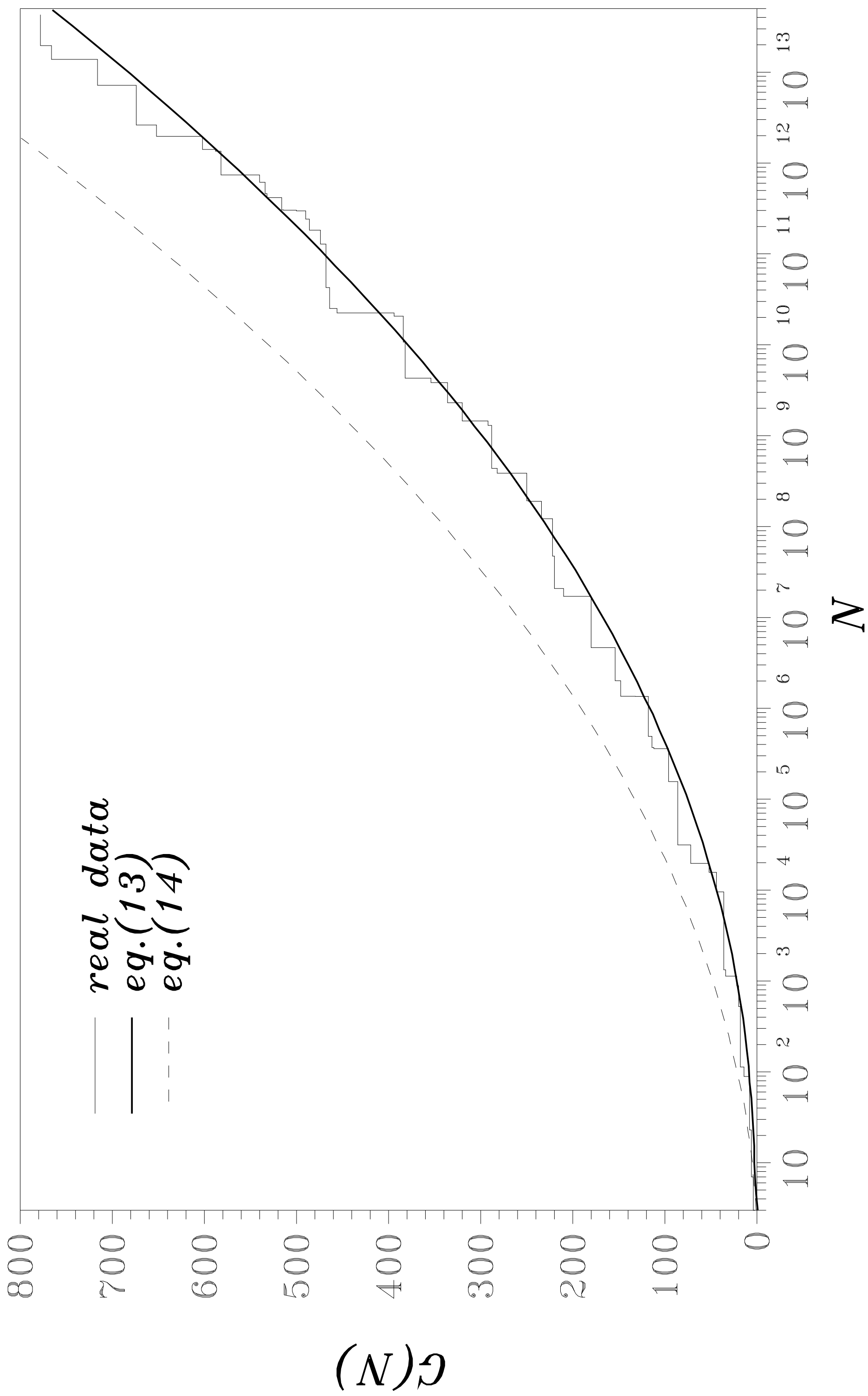

Fig.3 Plot of $G(N)$ for N up to 5×10^{13}. The results from the computer search are drawn by thin solid line, eq.(13) is bold line and the conjecture of Cramer (14) is shown by dashed line.

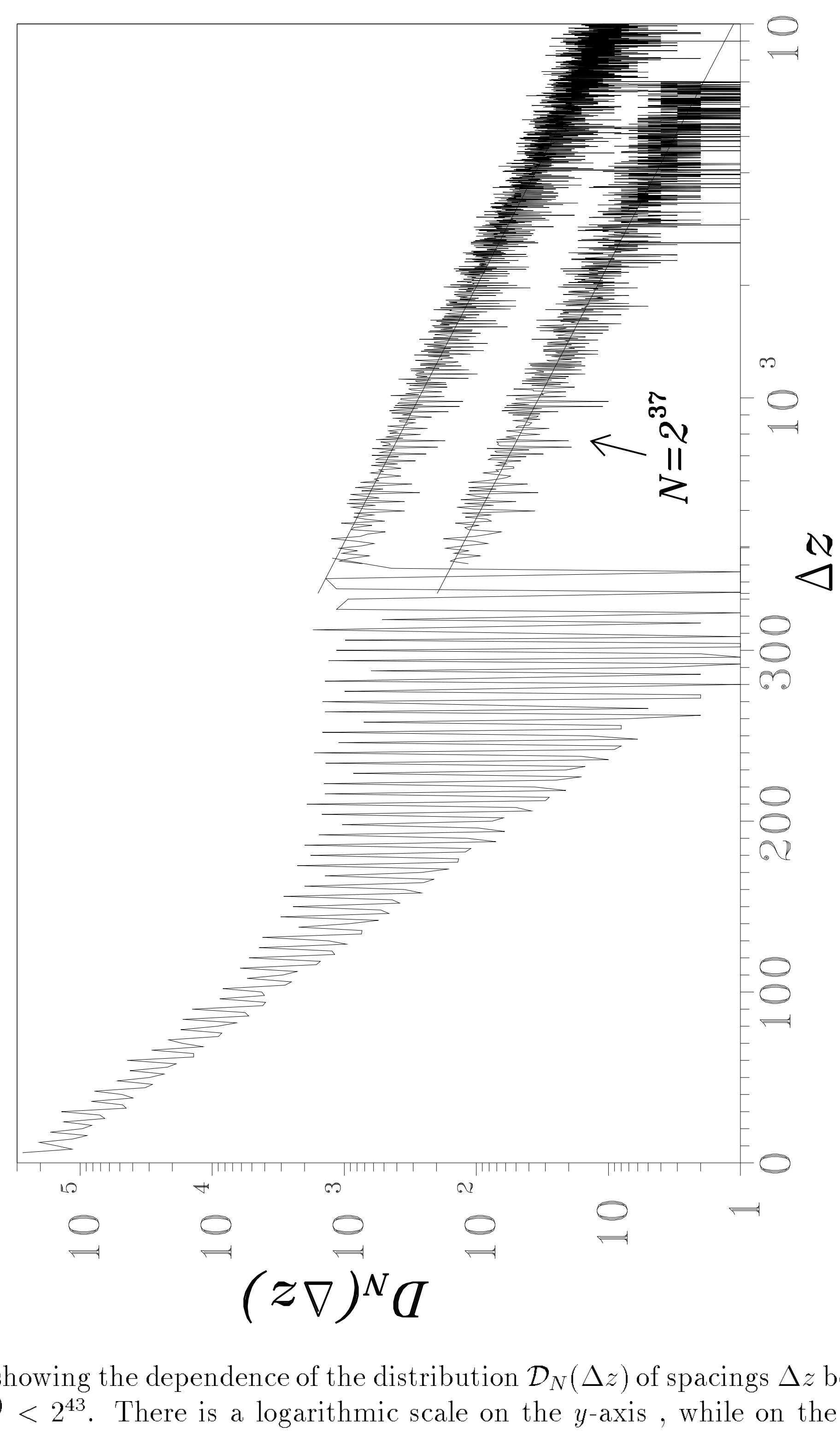

Fig.4 The plot showing the dependence of the distribution $\mathcal{D}_N(\Delta z)$ of spacings Δz between consecutive $p^{(z)} < 2^{43}$. There is a logarithmic scale on the y-axis , while on the x-axis there is a linear scale up to $\Delta z = 360$ while for larger Δz the scale is logarithmic. There is also the power-like part of $\mathcal{D}_N(\Delta z)$ for $N = 2^{37}$ shown and for this case the slope is $\gamma \approx 1.47$.