Some heuristics on the gaps between consecutive primes[1]

Marek Wolf

1. Introduction.

In 1922 Hardy and Littlewood [1] have proposed about 15 conjectures. The conjecture B of their paper states:

There are infinitely many primes pairs (p, p'), where $p' = p + d$, for every even d. If $\pi_d(N)$ denotes the number of pairs less than N, then

$$\pi_d(N) \sim 2c_2 \frac{N}{\log^2(N)} \prod_{p|d} \frac{p-1}{p-2}. \tag{1}$$

Here the constant c_2 is defined in the following way[2]:

$$c_2 \equiv \prod_{p>2} \left(1 - \frac{1}{(p-1)^2}\right) = 0.66016\ldots \tag{2}$$

The computer results of the search for pairs of primes separated by a distance d and smaller than N for $N = 2^{22}, 2^{24}, \ldots, 2^{40} \approx 1.1 \times 10^{12}$ are shown in the Fig.1 and they provide a firm support in favor of (1). The characteristic oscillating pattern of points is caused by the product

$$J(d) = \prod_{p|d, p>2} \frac{p-1}{p-2} \tag{3}$$

appearing in (1).

The formula (1) has not been proved yet, see however [2], [3], [4]. Even the particular case of $d = 2$ corresponding to the famous problem of existence of infinitely many twin primes is not solved and no progress is expected in the near future. Not much is known about the gaps between *consecutive* primes, what seems to be more interesting and important than the arbitrary pairs of primes treated by the Hardy–Littlewood conjecture. In this paper I will present simple heuristic reasoning leading to the formula for number of *consecutive* prime pairs smaller than a given bound N *expressed directly by* $\pi(N)$ — the total number of primes up to N. I will corroborate this hypothesis by showing that it leads to the well known formulas as the corollaries.

The problem of the appearance of gaps between consecutive primes has a long history. Two main questions related to that problem can be distinguished: One concerns the estimation of the difference:

$$d_n = p_{n+1} - p_n. \tag{4}$$

[1] revised in January 2006 version of the paper originally dated 1996

[2] The author believes there should exists heuristic explanation why the value of c_2 is so close to 2/3.

The growth rate of the form $d_n = \mathcal{O}(p_n^\theta)$ with different θ was proved in the past. The Riemann Hypothesis implies $\theta = \frac{1}{2} + \epsilon$ for any $\epsilon > 0$ and a few results with θ closest to 1/2 are: the result of Mozzochi [6] $\theta = \frac{1051}{1920}$, Lou and Yao obtained $\theta = 6/11$ [7] and recently Baker and Harman have improved it to $\theta = 0.535$ [8]. The second group of papers deals mainly with first appearance of a given gap of length d, see e.g. [9], [10], [11], [12] or the explicit formula giving the largest gap between consecutive primes [14], [15], [16], [17]. In 1974 there appeared the paper by Brent [18], where the statistical properties of the distribution of gaps between consecutive primes were studied both theoretically and numerically. Brent have applied the inclusion–exclusion principle and obtained from (1) the formula for the number of consecutive prime pairs less than N. But his result (formula (4) in [18]) had not a closed form and he had to produce on the computer the table of constants appearing in his formula (4). The attempt to estimate those sums and to write a closed formula for them was undertaken in [5]. However here I will present completely different approach to the problem of prime gaps.

2. The main conjecture

The above notation $\pi_d(N)$ denotes prime pairs not necessarily successive. In this paper I am interested in the investigation of a behaviour of the number of pairs of *consecutive* primes p_n, p_{n+1} with difference $p_{n+1} - p_n = d$. To make more clear distinction from the case treated by the Hardy–Littlewood conjecture, I exchange the appearance of index d and N in our notation:

$$h_N(d) = \text{number of pairs } \ p_n, p_{n+1} < N \text{ with } d = p_{n+1} - p_n. \tag{5}$$

Let me note that, skipping the only triple (3, 5, 7) of three consecutive primes separated by 2, the identities $\pi_d(N) \equiv h_N(d)$ hold for $d = 2$ (Twin primes) and $d = 4$, which are naturally to name "Cousin primes".

I have counted on a computer the number of gaps between consecutive primes up to $N = 2^{44} \approx 1.76 \times 10^{13}$. It took three CPU weeks on the DEC Alpha 500/500MHz to reach this bound. 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 $N = 2^{20}, 2^{22}, \ldots, 2^{42}, 2^{44}$. Such a choice of the intermediate thresholds as powers of 2 was determined by the employed computer program, because the primes were coded as bits. It took 3 CPU weeks on the DEC Alpha 500/500MHz workstation to reach 2^44. The resulting curves are plotted on the semi–logarithmic axes on the Fig.2. The straight lines are the least–square fits of the assumed exponential decrease of $h_N(d)$ with d to the actual values.

In the plot of $h_N(d)$ in the Fig.2 a lot of regularities can be observed. The pattern of points in Fig.2 *does not depend on* N: for each N the arrangements of circles is the same, only the intercept increases and the slope decreases. Comparison of the Fig.1 and Fig.2 and the fact that the points in Fig.2 lie around the straight lines on the semi-logarithmic scale suggest the following *Ansatz* for $h_N(d)$:

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

The essence point of the present consideration consists in a possibility of determining the unknown functions $A(N)$ and $B(N)$ by *assuming only the above exponential decrease of $h_N(d)$ with d and employing two identities fulfilled by $h_N(d)$*. First of all, the number of all gaps is by 1 smaller than the number of all primes smaller than N:

$$\sum_d h_N(d) \;=\; \pi(N) - 1, \tag{7}$$

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

$$\sum_d h_N(d)\; d \;\approx\; N. \tag{8}$$

The problem is that for the *Ansatz* (6) I am not able to write down in the closed form the selfconsistency conditions (7) and (8), because the product (3) behaves very erratically. However, $J(d)$ takes small values of the order 1; e.g. for $N < 2^{44}$ it takes values between 1 and 3.2 (3.2 appears only when $d = 210$, $d = 420$ or $d = 630$ — the largest gap for $N < 2^{44}$ was 716). To calculate the sums in (7) and (8) I replace the product $J(d)$ in (6) by its mean value s defined as:

$$s = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \prod_{p|2k, p>2} \frac{p-1}{p-2}. \tag{9}$$

Summing up geometrical series from the above identities (7), (8) and assuming that $A(N) \ll 1$ I get:

$$A(N) = \frac{\pi(N)}{N}, \tag{10}$$

$$B(N) = 2\frac{\pi^2(N)}{sN}. \tag{11}$$

Comparison with the Hardy and Littlewood conjecture for Twins $d = 2$ for $\log(N) \gg d$ gives $1/s = c_2$. In this heuristic way we arrive at the identity:

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \prod_{p|2k, p>2} \frac{p-1}{p-2} = \frac{1}{\prod_{p>2}(1 - \frac{1}{(p-1)^2})} = 1.514\ldots.$$

This equality indeed can be rigorously proved: it follows e.g. from the eq.(33) in [19] or eq. (17.12) in [20] in the limit $N \to \infty$. It can be also shown to be true by an application of the Ikehara-Delange theorem [21].

Finally we state the main conjecture:

$$h_N(d) = 2c_2 \frac{\pi^2(N)}{N} \prod_{p|d, p>2} \frac{p-1}{p-2} e^{-d\pi(N)/N} + error\ term(N, d) \;\text{ for } d \geq 6 \tag{12}$$

while for Twins ($d = 2$) and Cousins ($d = 4$) the identities $h_N(d) = \pi_d(N)$ hold. The formula (12) consists of three terms. The first one depends only on N, the second only on d, but the third term depends both on d and N.

Putting in (12) $\pi(N) \sim N/\log(N)$ and comparing with the original Hardy–Littlewood conjecture we have that the number $h_d(N)$ of *successive* primes $(p_{n+1},\ p_n)$ smaller than N and of the difference d $(= p_{n+1} - p_n)$ is diminished by the factor $\exp(-d/\log(N))$ in comparison with *all* pairs of primes (p, p') apart in the distance d $(= p' - p)$:

$$h_N(d) = \pi_d(N)e^{-d/\log(N)} + error\ term(N, d), \quad \text{for } d \geq 6. \tag{13}$$

Huristically this relation encodes in the series for $e^{-d/\log(N)}$ the inclusion-exclusion principle for obtaining $h_N(d)$ from $\pi_d(N)$. The above relation is illustrated by the Figures 1 and 2. Taking N very large at fixed d (such, that $\log(N) \gg d$) we get from (13) the intuitively obvious relation $h_N(d) \approx \pi_d(N)$ — the plots in the Fig.2 tend to the horizontal line for increasing N.

3. Two applications

As the application of the Conjecture 12 I 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.2:

$$h_N(G(N)) = 1. \tag{14}$$

Using the general form of the functions $A(N)$ and $B(N)$ (10) and (11) we obtain

$$G(N) \sim g(N) = \frac{N}{\pi(N)}(2\log(\pi(N)) - \log(N) + c_0), \tag{15}$$

where $c_0 = \log(2c_2) \approx 0.278$. The above prediction expresses $G(N)$ directly by $\pi(N)$ and for the estimation $\pi(N) \sim N/\log(N)$ conjectured by Gauss it gives:

$$G(N) \sim g_1(N) = \log(N)(\log(N) - 2\log\log(N) + c_0). \tag{16}$$

To my knowledge the above dependence of $G(N)$ is new, the most similar I have seen in [16], where Cadwell gave the heuristic arguments that

$$G(N) \sim \log(N)(\log(N) - \log\log(N)). \tag{17}$$

The formula (16) for large N passes onto the Cramer [14] conjecture (see also [15]):

$$G(N) \sim \log^2(N). \tag{18}$$

The examination of the formula (15) with the "experimental" results and the formula (18) is given in the Fig.3. Let me notice, that the values of $N \approx 10^{15}$ searched by direct checking are small for the asymptotic of eq.(16), because $\log 10^{15} = 36.84\ldots, \log\log 10^{15} = 3.61\ldots$, however large for modern computers, but nevertheless the agreement between the prediction and the actual data is quite good — there are about twenty sign changes of the difference $G(N) - g(N)$ on the Fig.3. This Figure should be compared with the figure on the page 12 in [23]. To produce this plot I have put in (15) the formula $\pi(N) \sim \int_2^N du/\log(u)$.

There are serious doubts about the validity of the Cramer conjecture [24]. A. Granville believes that the actual $G(N)$ can be larger than that given by (18), namely he claims [24] that there are infinitely many pairs of primes p_n, p_{n+1} for which:

$$p_{n+1} - p_n = G(p_n) > 2e^{-\gamma} \log^2(p_n) = 1.12292 \ldots \log^2(p_n). \tag{19}$$

However, by assuming the Hypothesis H of Sierpinski and Schinzel [25] it can be proved that for *almost* all prime gaps $G(N) < \log^2(N)$. In connection with this let me remark that on the Fig.3 the curve $G(N)$ always lies below the Cramer conjecture (18) (what agrees with the corollary from the Sierpinski–Schinzel Hypothesis) and for (19) to be true there should be such regions were $G(N) > 1.123 \log^2(N)$. In fact, the estimation (15) I have obtained using the formula for $h_N(d)$ outside the regime of its applicability — for large d the corresponding values of $h_N(d)$ are small and display large fluctuations (see Fig.2): $h_N(d)$ describes accurately only the "bulk" values of the number of gaps. There are large fluctuations between actual data and the formula (15) seen in the Fig.3, but there are also points, where the "staircase" like plot of $G(N)$ crosses the analytical curve $g(N)$.

The glance at the Fig.3 suggests that the following "average" property may hold true:

$$\lim_{N \to \infty} \frac{\int_2^N G(x)dx}{\int_2^N g(x)dx} = 1. \tag{20}$$

which could be in agreement with (19). The integral of the difference $G(N) - g(N)$ oscillates and only the ratio of surfaces below curves $G(N)$ and $g(N)$ has the chance to tend to 1.

TABLE **II**

The sum of squares of gaps between consecutive primes. In the second column the numbers obtained by a computer are given, while in the third one values obtained from eq.(21) and in the fifth from eq.(22) are presented. The fourth and sixth column contains the appropriate ratios.

N	$\sum d_n^2$	eq.(21)	ratio	eq.(22)	ratio
2^{20}	22171764	29072700	0.7626	26772796	0.8281
2^{22}	100275380	127919880	0.7839	118820844	0.8439
2^{24}	444929864	558195838	0.7971	522110155	0.8522
2^{26}	1959715564	2418848633	0.8102	2275494249	0.8612
2^{28}	8565851940	10419655651	0.8221	9849466113	0.8697
2^{30}	37168128504	44655667077	0.8323	42385671111	0.8769
2^{32}	160316134724	190530846196	0.8414	181487345070	0.8833
2^{34}	687851546612	809756096335	0.8495	773707462336	0.8890
2^{36}	2938092559092	3429555231536	0.8567	3285796459875	0.8942
2^{38}	12499933597196	14480344310930	0.8632	13906838608376	0.8988
2^{40}	52993288896472	60969870782865	0.8692	58681260682528	0.9031
2^{42}	223959886541176	256073457288032	0.8746	246938313792889	0.9069
2^{44}	943825347126668	1073069725778421	0.8796	1036598392711419	0.9105

As the final application of the formula (12) and our computer data we consider the conjecture made by D.R. Heath-Brown in [26], see also problem A8 in [27]. Namely

Heath-Brown guessed that

$$\sum_{p_n \leq N} (p_n - p_{n-1})^2 \sim 2N \log(N). \tag{21}$$

Treating the product $J(d)$ exactly the same way as in derivation of (10) and (11) from (12) I get:

$$\sum_{p_n < N} (p_n - p_{n-1})^2 = \sum_d d^2 h_N(d) = \frac{2N^2}{\pi(N)} + error\ term(N). \tag{22}$$

Again, the above formula involves directly $\pi(N)$ and for $\pi(N) \sim N/\log(N)$ passes over onto the (21). The comparison with the computer data is given in Table II and it suggests that (22) better reproduces data obtained by computer than (21), however ratios of both predictions to the actual computer data tend to 1 with increasing N.

Acknowledgment: I thank Prof. E. Bombieri and Prof. A.Granville for e-mail feedback and Prof. W.Narkiewicz for discussions and Prof. T. Nicely for sending me new maximal prime gaps.

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
home page: www.ift.uni.wroc.pl/∼mwolf

References

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

[2] Lord Cherwell, "Note on the Distribution of the intervals between prime numbers", Quart. J. of Math. (Oxford) **17** (1946) 46–62

[3] G. Polya, "Heuristic Reasoning in the Theory of Numbers", *American Math. Monthly*, **66** (1959), 375–384

[4] M.Rubinstein "A Simple Heuristic Proof of Hardy and Littlewood's Conjecture B", *American Math. Monthly*, **100** (1993), 456–460

[5] A. Odlyzko, M.Rubinstein, M.Wolf, *Jumping Champions*, *Experimental Mathematics* **8** (1999), p.107

[6] C.J. Mozzochi, "On the difference between consecutive primes", *Journal Number Theory*, **24** (1986), p. 181–187

[7] S.Lou and Q.Yao, "A Chebyshev's type of prime number theorem in a short interval. II.", *Hardy-Ramanujan J.* **15**, (1992), p. 1-33

[8] Baker, R.C. and Harman, G. "The difference between consecutive primes", *Proc. Lond. Math. Soc.*, III. Ser. 72, No.2 (1996), p.261–280

[9] L.J.Lander and T.R.Parkin, "On First Apperance of Prime Differences", *Math.Comp.* **21** (1967), 483

[10] R.P.Brent, "The First Occurrence of Certain Large Prime Gaps", *Math. Comp.***35** (1980), 1435-1436

[11] J.Young and A.Potler, "First occurrence Prime Gaps", *Math.Comp.* **52** (1989), 221-224

[12] T.Nicely, "New Maximal Prime Gaps and First Occurrences", *Math. Comp.* **68** (1999), 1311-1315;

[13] T.Nicely and B. Nyman, "First occurrence of a prime gap of 1000 or greater", preprint available at http://purl.oclc.org/NET/TRN

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

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

[16] J.H.Cadwell, "Large Intervals Between Consecutive Primes", *Math.Comp.***25** (1971), 909

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

[18] R.P.Brent, "The Distribution of Small Gaps Between Successive Primes", *Math.Comp.***28** (1974), 315-324

[19] E. Bombieri and H. Davenport, "Small differences between prime numbers", *Proc. Royal Soc.*, **A293**, (1966), 1-18

[20] H.L. Montgomery, "Topics in Multiplicative Number Theory", Springer Lecture Notes 227 (Heidelberg, New York, 1971)

[21] W. Narkiewicz, private communication

[22] R.P.Brent, "Irregularities in the Distribution of Primes and Twin Primes", *Math.Comp.* **29** (1975), 43-56

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

[24] A.Granville,"Unexpected irregularities in the Distribution of Prime Numbers", in Proceedings of the International Congress of Mathematicians (Zurich, Switzerland, 1994), Vol. I (1995), 388–399.

[25] A.Schinzel, "Remarks on the paper "Sur certain hypotheses conernant les nombres premiers"", *Acta Arithmetica* vol. VII (1961), p.1–8

[26] D.R.Heath-Brown, "Gaps between primes, and the pair correlation of zeros of the zeta-function", *Acta Arithmetica* vol.XLI (1982), 85

[27] R.K. Guy, *Unsolved Problems in Number Theory*, (Springer–Verlag, New York, Heidelberg, 1994)

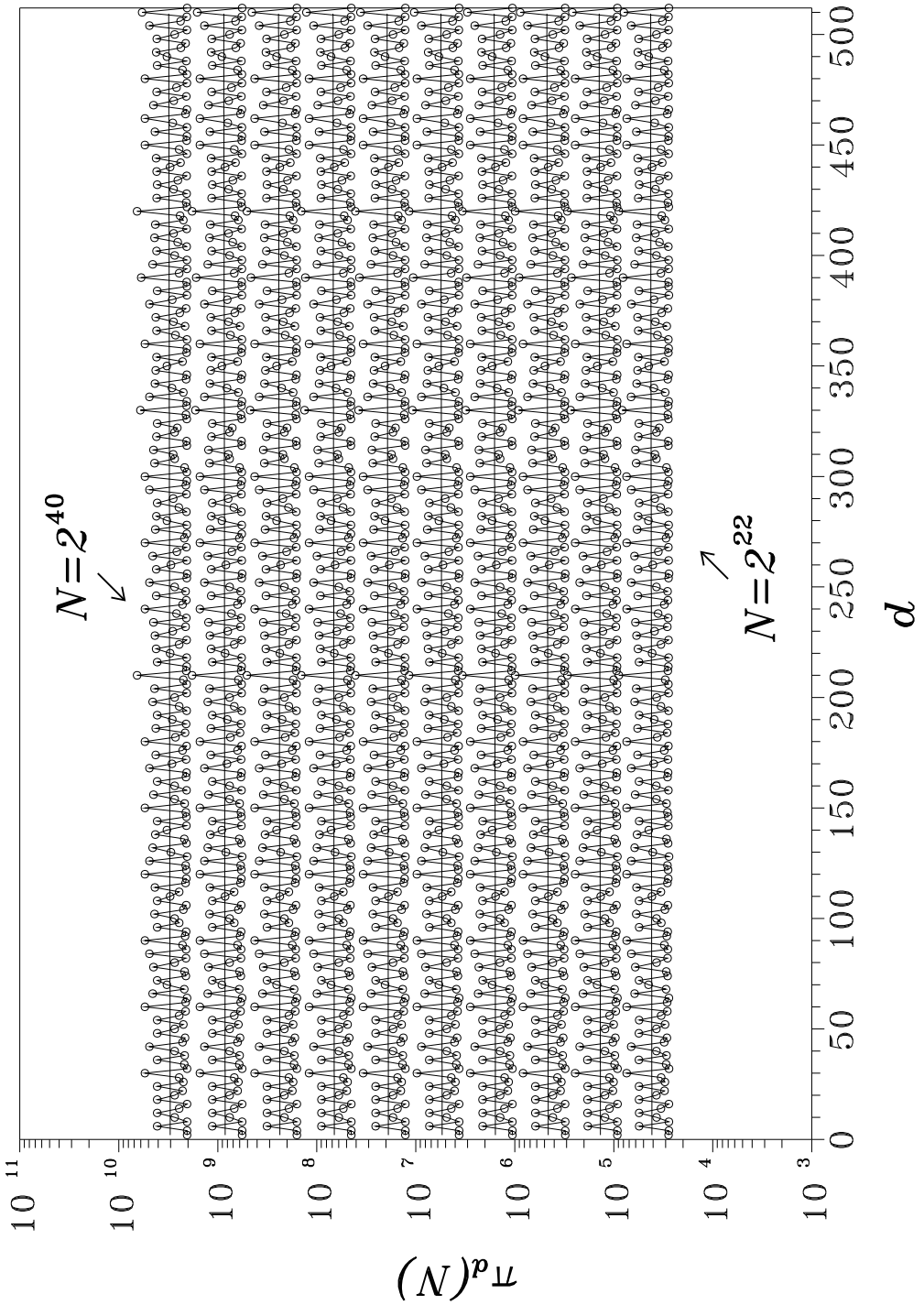

Figure 1: The plot illustrating the Hardy–Littlewood conjecture. There is a logarithmical scale on the y-axis , while on the N-axis there is a linear scale. There are oscillations of the period 6 clearly visible. Additionally, there are structures of the length 30 overimposed. Let us mention "local" spikes at $d = 30, 60, \ldots$ and especially well profound spikes at $d = 210 = 2 \cdot 3 \cdot 5 \cdot 7$ — they are called "jumping champions", see [5].

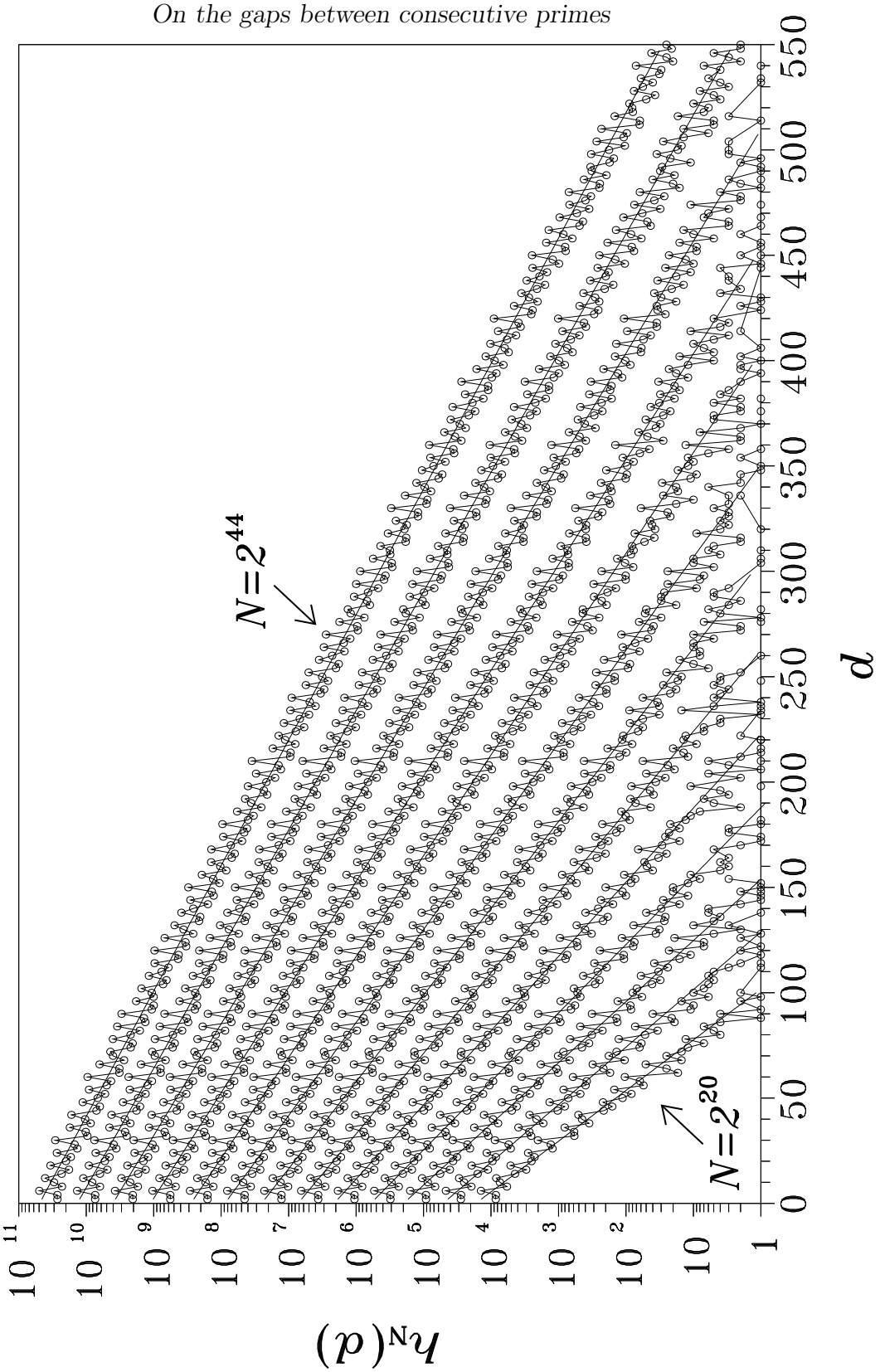

Figure 2: The plot showing the dependence of the histogram $h_N(d)$ on d at $N = 2^{20}, 2^{22}, \ldots, 2^{44}$. There is a logarithmical scale on the y-axis , while on the x-axis there is a linear scale. The values of $h_N(d)$ obtained from the computer search are represented by small circles. The 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.

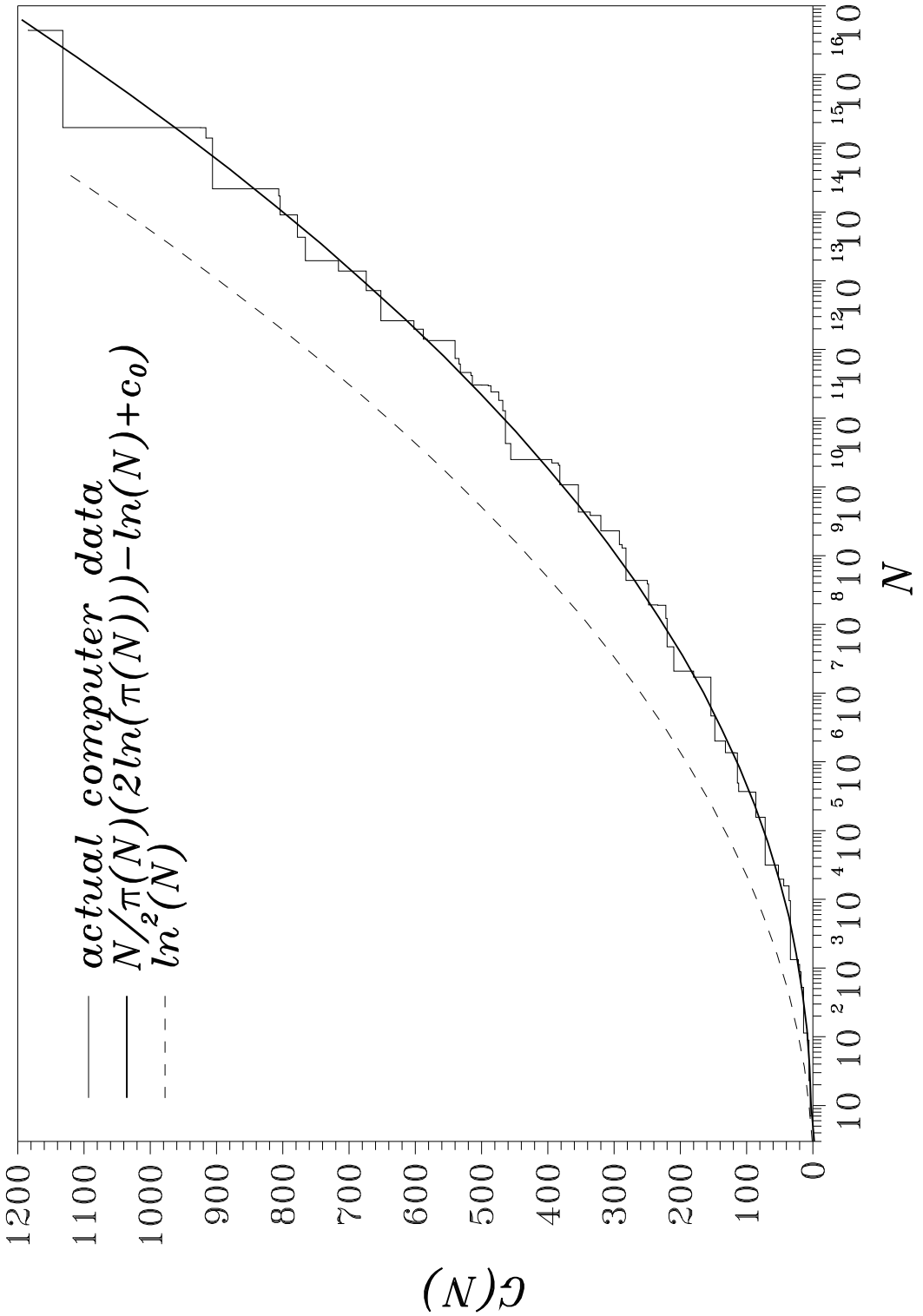

Figure 3: Plot of $G(N)$ for N up to 1.35×10^{15}. The results from the computer search are drawn by thin solid line, eq.(15) is bold line and the conjecture of Shanks (18) is shown by dashed line. Most of the points plotted on this figure come from my own search up to $2^{44} = 1.76 \times 10^{13}$. Gaps larger than 2^{44} I have taken from [11], [12] and [13].