ELSEVIER

Physica A 250 (1998) 335–344

PHYSICA A

Random walk on the prime numbers

Marek Wolf*

Institute of Theoretical Physics, University of Wrocław, Pl. Maxa Borna 9, PL-50-204 Wrocław, Poland

Received 3 July 1997

Abstract

The one-dimensional random walk (RW), where steps up and down are performed according to the occurrence of special primes, is defined. Some quantities characterizing RW are investigated. The mean fluctuation function $F(l)$ displays perfect power-law dependence $F(l) \sim l^{1/2}$ indicating that the defined RW is not correlated. The number of returns of this special RW to the origin is investigated. It turns out that this *single*, very special, realization of RW is a typical one in the sense that the usual characteristics used to measure RW, take values close to the ones averaged over *all* random walks. This fact suggests that random numbers of good quality could be obtained by means of RW on prime numbers. The fractal structure on the subset of primes is also found. © 1998 Elsevier Science B.V. All rights reserved.

Keywords: Prime numbers; Random walks; Fractals; Random number generators

In 1973 Bilingsley published the paper "Prime Numbers and Brownian Motion" [1], where he defined the set of random walks (RW) employing the factorization of integers into the primes. In this paper we are going to define even simpler RW on the primes, which uses special families of primes. Namely, among the primes the subset of Twin primes is distinguished: Twins are such numbers $\{p, p'\}$ that both p and $p' = p+2$ are prime. So the set of Twins starts with (3, 5), (5, 7), (11, 13), (17, 19), (29, 31),... . It is not known whether there is an infinity of Twins; the largest pair of Twins known today was found recently by Indlekofer and Jarai [2]:

$$697\,053\,813 \times 2^{16352} \pm 1\,. \qquad (1)$$

Let us notice that these same authors have announced on the Internet the much larger pair of Twins: $242\,206\,083 \times 2^{38880} \pm 1$.

The mathematicians are using the notation $\pi_2(N)$ to denote the number of Twins smaller than N. The next possible gap (after 2) between consecutive primes is 4 and we

* E-mail: mwolf@ift.uni.wroc.pl.

PII S0378-4371(97)00661-4

Table 1
The numbers of Twins ($d=2$) and Cousin ($d=4$) up to $N=2^{18}$, $N=2^{20}, \dots N=2^{44}$

N	$\pi_2(N)$	$\pi_4(N)$	$\pi_2(N)/\pi_4(N)$
2^{18}	2679	2678	1·00037
2^{20}	8535	8500	1.00412
2^{22}	27 995	27 764	1.00832
2^{24}	92 246	91 995	1.00273
2^{26}	309 561	309 293	1.00087
2^{28}	1 056 281	1 057 146	0.99918
2^{30}	3 650 557	3 650 515	1.00001
2^{32}	12 739 574	12 740 283	0.99994
2^{34}	44 849 427	44 842 399	1.00016
2^{36}	159 082 253	159 089 620	0.99995
2^{38}	568 237 005	568 225 073	1.00002
2^{40}	2 042 054 332	2 042 077 653	0.99999
2^{42}	7 378 928 530	7 378 989 766	0.99999
2^{44}	26 795 709 320	26 795 628 686	1.00000

will use the name Cousins to denote such numbers $\{p, p'\}$ that both p and $p'=p+4$ are prime. Examples of Cousins are (7, 11), (13, 17), (37, 41),… . The function $\pi_4(N)$ will denote the number of Cousins smaller than N. The pairs of primes separated by $d=2$ and $d=4$ are special as they always have to be consecutive primes, with the exception of the pair (3,7) containing 5 in the middle – for primes $p, p+2k$ separated by gaps with $k \geqslant 3$ there is a possibility to have primes in between. For example, for $k=3$ it is possible to have between p, $p+6$ the prime either of the form $p+2$ or $p+4$: triplets (7, 11, 13) or (41, 43, 47) serve as examples.

In Ref. [3] Hardy and Littlewood have conjectured, that the number of Twins and Cousins below a given bound N should be approximately equal to each other and given by the approximate formula:

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

where $d=2,4$. Here the constant c_2 (sometimes called "twin-prime" constant) is defined in the following way:

$$c_2 \equiv 2 \prod_{p>2} \left(1 - \frac{1}{(p-1)^2}\right) = 1.32032\dots . \tag{3}$$

But it turns out that the computer search up to $N=2^{44}$ shows that the relation

$$\pi_2(N) \approx \pi_4(N) \tag{4}$$

is well satisfied already for small values of N – it holds not necessarily asymptotically for large N. Table 1 shows the values of the numbers of Twins and Cousin captured during the computer scan at the values of N forming the geometrical progression $N=2^{18}, 2^{20}, \dots, 2^{44}$. Because Twins and Cousin seems to appear randomly and the number of occurrences of them is almost the same, we can define the one-dimensional random walk in the following way: Let us move along consecutive integers

$1, 2, 3, 4, \dots$. If we meet the pair of Twins, then the random walker makes the step say up and if the pair of Cousin is encountered, then the step down is performed. I will use the abbreviation PRW to denote this special random walk on the primes. The similarity with the random walk defined by the structure of DNA sequences [4] can be mentioned at this point. If there is infinity of Twins and Cousin (as suggested by the Hardy – Littlewood conjecture), then the PRW defined above will continue to perform steps forever, in contrast to RW considered in Ref. [1] or Ref. [4], where random walks were finite.

Let $y(N)$ denote the displacement of the random walker after N steps, hence $y(N)$ represents the function:

$$y(N) = \pi_2(N) - \pi_4(N) . \tag{5}$$

The function $y(N)$ is piecewise constant, because $\pi_d(N)$ can change values only for N being the prime. The Fig. 1 shows the plot of $y(N)$ in the range $N \in (1, 10^{12})$. The argument is shown on the logarithmical N-axis. The comparison of three parts (a), (b) and (c) of Fig. 1 reveals a self-affinity of $y(N)$: left parts starts with oscillations of relatively small amplitudes and they increase at right parts (Figs. 1a–c have increasing scales on the y-axis). Let me remark that, if instead of using the logarithmic scale, the x-axis would be drawn linearly with the same yardstick as the vertical axis on the original of Fig. 1c, where the interval (0, 25 000) had the length 11 cm, then the total plot for $1 \leqslant N \leqslant 10^{12}$ should be 4400 km long! It took over 9 days of CPU time on the DEC 3000/800 200 MHz workstation to produce the data for Fig. 1.

The main question on the defined above PRW is whether the consecutive steps are correlated or not. In other words: do the appearance of a Twin or Cousin depend on the previous history? The standard answer to this question is obtained by calculation of the mean square fluctuation $F(l)$ about the average of the displacement, [4,5]. The function $F(l)$ is defined by the equation:

$$F^2(l) = \langle (\Delta y(l))^2 \rangle - \langle \Delta(l) \rangle^2 , \tag{6}$$

where $\Delta y(l) = y(l + l_0) - y(l_0)$ and the average is performed over all starting points l_0 in the random walk. For the usual random walk the function $F(l)$ is described by the power law:

$$F(l) \sim l^{\alpha} . \tag{7}$$

with the exponent $\alpha = \frac{1}{2}$. The exponents $\alpha \neq \frac{1}{2}$ characterize the walks that display long-range correlations between consecutive steps: positive for $\alpha > \frac{1}{2}$ (it means that if, say step up was performed, then it is more likely to perform again the step up), and negative for $\alpha < \frac{1}{2}$, see e.g. Ref. [5]. This function $F(l)$ was used in the past to reveal the long-range correlations in the DNA sequences [4].

In principle, the relation of the form (7) holds in the limit of infinitely long walks, but in practice only finite-step RW are observed. We have generated the PRW up to $N = 2^{39} \approx 5.5 \times 10^{11}$. Because PRW performs steps only at Twins or Cousins, the values of l_0 were not consecutive integers, but were chosen as multiplicities of 64 (it was

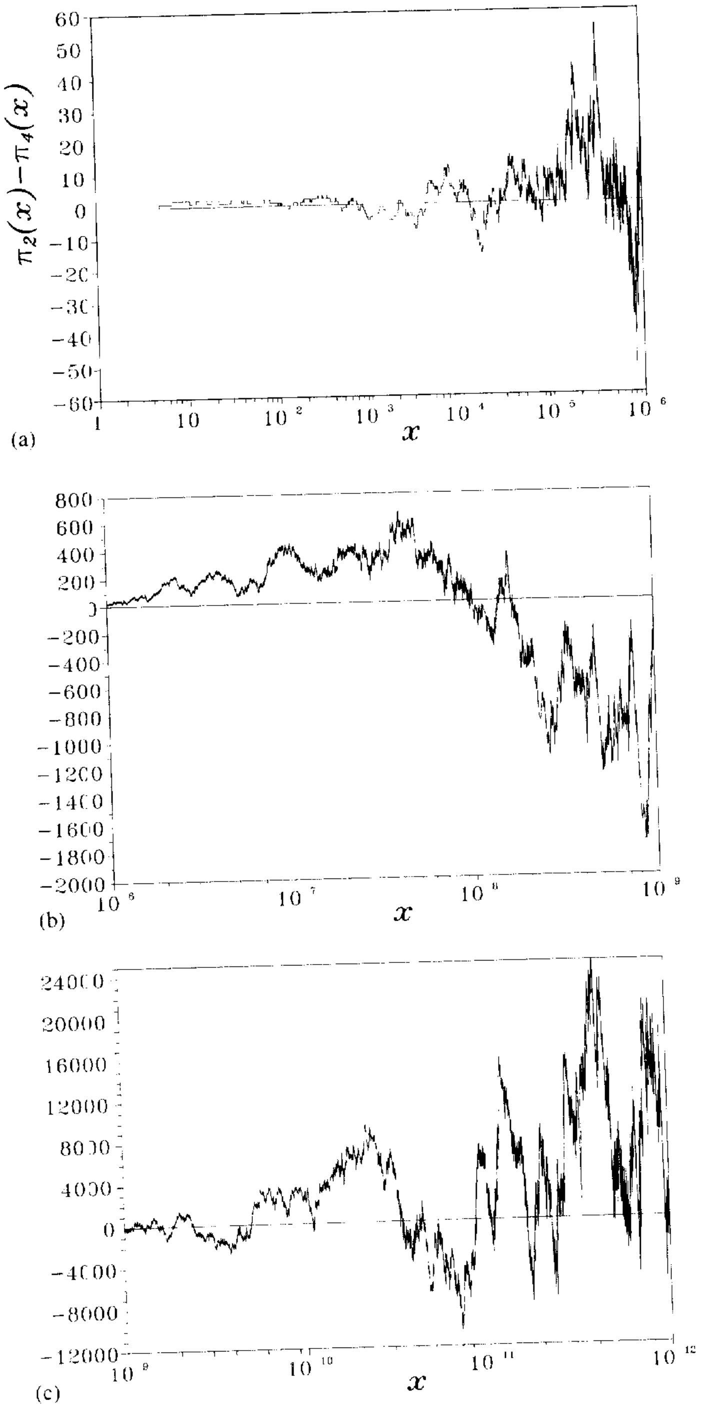

Fig. 1. The function $y(N)$ is plotted in the range $N \in (1, 10^6)$ in part (a), for $N \in (10^6, 10^9)$ in (b) and for $N \in (10^9, 10^{12})$ in (c). The range of values on the y-axis changes in each case. Up to $N = 5 \times 10^6$ all arguments N are plotted – for arguments larger some decimation procedure was employed. Namely, for $y(N) > 100$ only changes of values larger than 8% were recorded, while smaller values of the function $y(N)$ were updated only for changes larger than 30%. This procedure causes that there are approximately 31 000 points in part (a), 150 000 points in part (b) and 250 000 points in part (c), while the total number of Twins and Cousins was in fact much larger (see Table 1). Let me remark that if the x-axis would be drawn with the same yardstick as the vertical axis, then it should be 4400 km long.

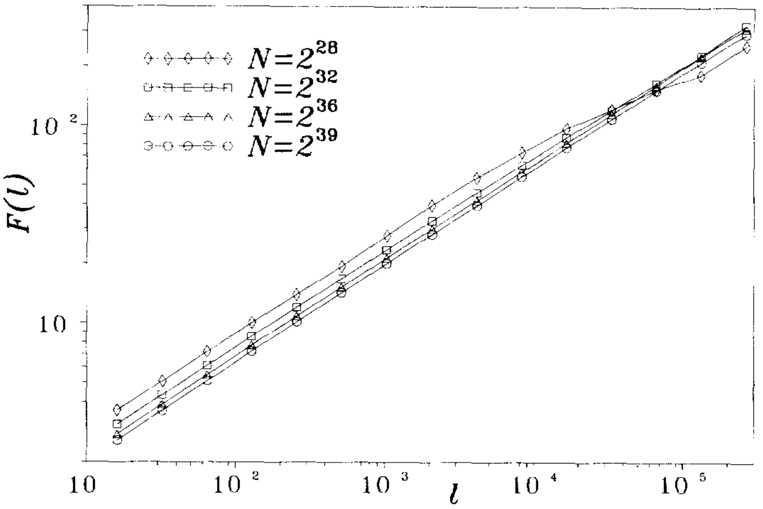

Fig. 2. The plot of finite approximations to $F(l)$ for $N=2^{28}$, 2^{32}, 2^{36}, 2^{39}. The curves converge for increasing N to the asymptotic plot for $N=\infty$.

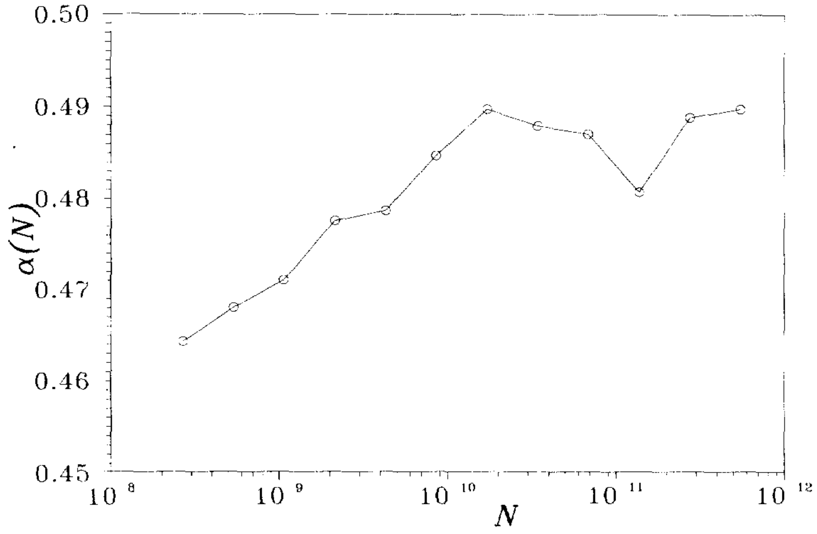

Fig. 3. The plot of the slope $\alpha(N)$ calculated from the partial functions $F(l)$.

caused by the 64-bit architecture of DEC Alpha processor on which all programs were run). The function $F(l)$ was calculated at discrete set of lengths $l=16,32,64,\ldots,2^{18}=262\,144$. The partial approximations of the function $F(l)$ were stored at $N=2^{28},2^{29},\ldots,2^{39}$ (the number of different l_0 used for averaging was $N/64$). The plot of the $F(l)$ for a sample of N is plotted on the double logarithmic scale in Fig. 2. The partial approximates converge with increasing N to the limiting curve. The dependence of the exponent $\alpha(N)$ on the number of steps N is shown in the Fig. 3. After the initial increase from 0.46 for $N=2^{28}$ to 0.49 for $N=2^{35}$ the exponent $\alpha(N)$ begins to fluctuate

Table 2
The values of the function $\pi_z(N)$ giving the number of primes $p^{(z)} < N$ fulfilling $y(p^{(z)}) = 0$

N	$\pi_z(N)$
1000	31
10 000	60
100 000	592
1 000 000	2332
10 000 000	2332
100 000 000	4718
1 000 000 000	15 351
10 000 000 000	68 440
100 000 000 000	278 503
1 000 000 000 000	1 787 793
10 000 000 000 000	2 823 290

around the value 0.49. Such very close to 0.5 values of α indicate that PRW does not display correlations.

Another quantity characterizing RW is the number of returns to the origin see e.g. Ref. [6]. The returns to the origin happens when $y(N) = 0$ and let $\mathcal{T}(N)$ denote the set of such primes $p^{(z)}$ at which the numbers of Twins and Cousin are the same:

$$\mathcal{T}(N) = \{p^{(z)} < N : y(p^{(z)}) = 0\} . \tag{8}$$

The direct computer search shows that up to $N = 2^{43} \approx 8.8 \times 10^{12}$ there are 2 823 290 such primes $p^{(z)}$ that $y(p^{(z)}) = 0$ holds. First the same number of Twins and Cousin appears between 101 and 103 (besides the trivial zeros 2 and 3, when $\pi_2(N) = 0$ and $\pi_4(N) = 0$). The largest captured zero below 2^{43} was $8\,205\,034\,088\,567 \approx 2^{42.899646}$. On the Fig. 1a there are 2334 zeros, on (b) $y(N)$ touches x-axis 13019 times and in (c) there were 1 035 496 prime zeros of $y(N)$.

Let

$$\pi_z(N) = \{\textit{number of } p^{(z)} < N \textit{ such that } y(p^{(z)}) = 0\} \tag{9}$$

denote the number of returns to the origin of PRW defined by Eq. (5). In Table 2 the numbers of primes $p^{(z)}$ up to 10^{13} every one order of magnitude are given. The values of $\pi_z(N)$ in this table display rather large fluctuations however using the analogy with random walk one can expect that the $\pi_z(N)$ is of the same order as the number of visits of usual random walk to the origin. It is well known, see e.g. Ref. [6], that in one dimension the average number of returns of the random walk to the origin during n steps is $\sqrt{n/\pi}$ and we can expect that

$$\pi_z(N) \sim \sqrt{N/\pi} . \tag{10}$$

The comparison of this formula with the actual data is provided in Fig. 4. Taking into account the fact that theoretical prediction $\sqrt{n/\pi}$ holds for the number of returns of RW to the origin *averaged* over many samples, and the PRW defined on the primes

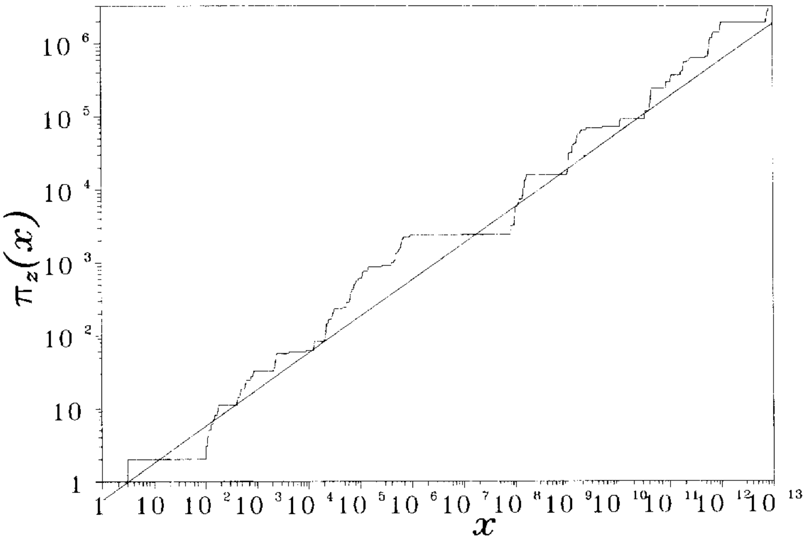

Fig. 4. The plot of the number of returns to the origin of the PRW. The straight line represents the plot of $\sqrt{N/\pi}$ – the conjectured dependence of the $\pi_z(N)$.

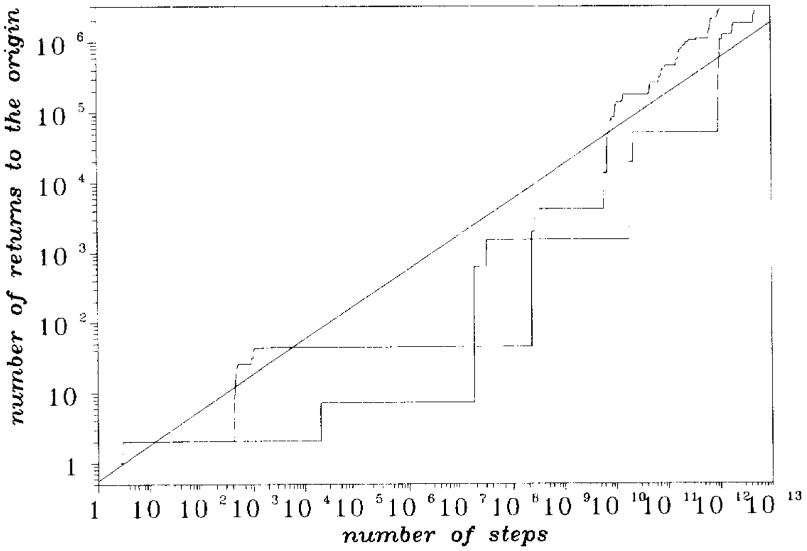

Fig. 5. The plot showing the dependence of the number of returns to the origin of the usual RW of size comparable to the number of returns in Fig. 4.

by Eq. (5) is only *one*, *particular* realization of RW, the agreement seems to be quite well. For the comparison in the Fig. 5, we have plotted the numbers of returns to the origin of two realizations of the usual RW in one-dimension. The simulation was stopped when the number of returns to the origin was comparable to the number of $p^{(z)}$ for the case of PRW: the two plots in Fig. 5 represent 2 826 062 and 2 770 354 returns to the origin and they consisted of over 4.9×10^{12} and 1.06×10^{12} steps respectively. To generate RW of length $\sim 10^{12}$ steps in reasonable time we have used the random bits generators based on primitive polynomials modulo 2, see Ref. [7].

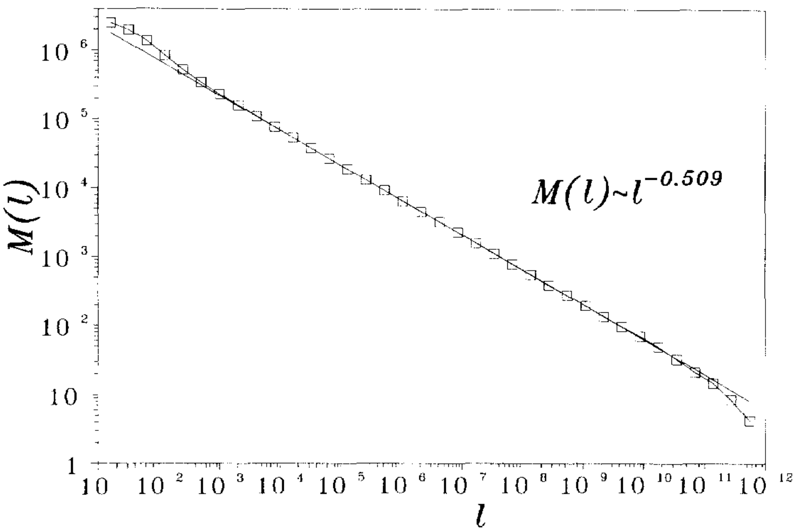

Fig. 6. The plot showing the dependence of the number of boxes $M(l)$ as a function of length l for $l=16,32,64,\dots,2^{39}\approx 5.5\times 10^{11}$. There is a logarithmic scale on the x and y-axes and the slope gives the fractal dimension of the set $\mathcal{T}(N)$.

The set $\mathcal{T}(N)$ consists of primes of two kinds organized in the clusters like structure. The "cluster" starts at the prime p' belonging to the pair of Twins or Cousin when the equality $\pi_2(p')=\pi_4(p')$ begins to hold. Next primes p not being Twins or Cousin are met and the equality $y(N)=0$ is maintained: inside "clusters" of p, the values of $\pi_2(N)$ and $\pi_4(N)$ do not change their values and are equal to each other. When there appears a prime belonging to the Twin or Cousin pair, then $\pi_2(N)$ or $\pi_4(N)$ increases by 1 and the equality $y(N)=0$ is lost. In other words, the first elements of these clusters are characterized by the equation:

$$y(p')=0\wedge y(p'-\varepsilon)\neq 0 \tag{11}$$

and the ends of clusters satisfy

$$y(p')\neq 0\wedge y(p'-\varepsilon)=0\,, \tag{12}$$

where $\varepsilon>0$. It turns out that the "clusters" (or "islands") of such N that $\pi_2(N)=\pi_4(N)$ are organized in a hierarchical, selfsimilar set.

To show it let us calculate the fractal dimension [8] of the set $\{p^{(z)}\}$. We have used the direct box-counting method. Namely, the whole interval $(1,2^{43})$ was covered by consecutive intervals of length $l=16$ and the number $M(l)$ of "boxes" containing primes $p^{(z)}$ was calculated. This procedure was successively repeated for lengths of boxes two times larger, up to $l=2^{39}\approx 5.5\times 10^{11}$. The obtained values of $M(l)$ are plotted in Fig. 6 in the double logarithmic axes. The large linear part in the middle tells us that

$$M(l)\sim l^{-D_{fr}},\quad D_{fr}\approx 0.509\,, \tag{13}$$

where D_{fr} is the fractal dimension of $\mathscr{T}(N)$, see e.g. Ref. [8]. It is well known, see Ref. [8] (p. 390) or Ref. [9], that the Haussdorf measure of the zeroset of the usual Brownian motion is equal to $\frac{1}{2}$, so the above value $D_{fr} = 0.509$ of Eq. (13) is very close to the strict, analytical prediction. There is a surplus of small boxes caused by the cluster–like organization of the set $\mathscr{T}(N)$: short boxes grasp separate primes at which $y(N)$ is not changing value (equal 0). In other words, there is a minimal length (depending on N), below which $\mathscr{T}(N)$ is not a fractal: the small boxes intersect with zeros of $y(N)$ which are inside "clusters". The fractal, selfsimilar hierarchy is formed only by primes $p^{(z)}$ marking the beginnings or ends of clusters and those are distinguished by conditions (11) and (12), respectively. The size of the largest cluster encountered during the computer search is not known, but the fact that boxes with $l > 1024$ follow perfectly the power-like dependence (they contain clusters totally inside) suggests that the length of clusters was smaller than 1000.

The fact, that usual characteristics of PRW are so close to the theoretical values obtained after averaging over *all* realizations suggests that the difference between Twins and Cousin can be used as a very fast random number generator of bits. For example, our program using Eratosthenes sieve was able to search on the DEC Alpha 200 MHz workstation for Twins and Cousins up to $2^{32} \approx 4.3 \times 10^9$ in less than 9 min. During that time random walk of over 25 millions *uncorrelated* steps are generated, see Table 1. Random bits can be mapped into the byte representation of floating numbers and used to generate usual RND numbers. Because at least 32 bits are needed to obtain one floating number (in single precision), the speed of such RND generators is of interest. However, it is known [7] that routines generating uncorrelated RND and simultaneously possessing long periods are rather complicated and hence time consuming, see e.g. RAN2 and RAN3 in Ref. [7], thus it is possible that RND based on PRW could be competetive in this respect. In addition, recently it was discovered that even RAN3 possesses some drawbacks [10,11]. Besides that, if there is an infinity of Twins and Cousins, as every mathematician believes, a random number generator based on the difference of Twins and Cousins can have *infinite period*. To be more precise, infinite period will emerge if there is no repeating pattern in the distribution of sequences of consecutive Twins and Cousins – infinitude of Twins and Cousins alone is not sufficient. But such a periodicity in the distribution of Twins and Cousins seems to be more curious than the absence of it. The possibility of using primes for generation of RND is now under study.

The author receives support from the KBN research grant No. 2 P302 057 07.

References

[1] P. Billingsley, Prime numbers and Brownian motion, Amer. Math. Monthly 80 (1973) 1099–1115.

[2] K.H. Indlekofer, A. Jarai, Math. Comp. 65 (1996) 427–428.

[3] G.H. Hardy, J.E. Littlewood, Acta Math. 44 (1922) 1–70.

[4] C.-K. Peng, S.V. Buldyrev, A.L. Goldberger, S. Havlin, F. Sciortino, M. Simons, H.E. Stanley, Nature 356 (1992) 168.

[5] A.A. Tsonis, J.B. Elsner, J. Stat. Phys. 81 (1995) 869–880.

[6] W. Feller, An Introduction to the Probability Theory and its Applications, vol. 1, Wiley, New York, 1970.

[7] W.H. Press, B.P. Flannery, S.A. Teukolsky, W.T. Vetterling, Numerical Recipes. The Art of Scientific Computing, Cambridge University Press, Cambridge, 1989.

[8] B.B. Mandelbrot, The Fractal Geometry of Nature, Freeman, San Francisco, 1982.

[9] E. Perkins, Zeit. Wahrsch. verw. Gebiete 58 (1981) 373–388.

[10] I. Vattulainen, T. Ala-Nissila, K. Kankaala, Phys. Rev. Lett. 73 (1994) 2513.

[11] I. Vattulainen, Doc:oral thesis, 1994, available at URL http://www.physics.helsinki.fi/tft/tft_preprints94.html.