# Random Sieving and the Prime Number Theorem

Karl Greger

*Karl Greger is Chairman of the Mathematics Department of the High School of Education in Gothenburg, Sweden. He has published several books, and his chief interests are theory of numbers, probability theory and applied and computer-oriented mathematics in general.*

## 1. INTRODUCTION

The Prime Number Theorem is a remarkable and rather deep result in the theory of numbers. The aim of this article is to show that this theorem can be made plausible by quite simple and elementary methods which, in addition, give a fascinating insight into the stochastic nature of that theorem.

The material of this article is intended for use at the college level to integrate and motivate chapters on probability, sequence and series (especially the harmonic series), and the logarithmic function usually treated in most elementary mathematics courses.

The heuristic arguments in this article can be tightened up, but that cannot be done at college level.

For every real number $x$ let $\pi(x)$ be the number of primes less than or equal to $x$. One finds that $\pi(10) = 4$, $\pi(100) = 25$, $\pi(1000) = 168$, etc. The function $x \to \pi(x)$ will be called the *prime number function*. All attempts to find a formula for $\pi(x)$ representing $\pi(x)$ in "closed form" by a finite number of "known" functions have failed, and will necessarily fail. There are, however, some simple asymptotic expressions for $\pi(x)$, such as

$$\pi(x) \sim \int_2^x \frac{dt}{\log t} \sim \frac{x}{\log x}.$$

Two expressions $f(x)$ and $g(x)$ are said to be asymptotically equal, $f(x) \sim g(x)$, if $\lim_{x\to\infty} f(x)/g(x) = 1$.

The fact that $\int_2^x (1/\log t)\, dt$ and $x/\log x$ are asymptotically equal is easily established by partial integration, while the fact that

$$\pi(x) \sim \int_2^x \frac{dt}{\log t} \qquad (1)$$

is a rather deep result in the theory of numbers, usually called the Prime Number Theorem. The following table shows that the integral logarithm $\int_2^x (1/\log t)dt$ is an excellent approximation to $\pi(x)$:

| $x$ | $\pi(x)$ | $\int_2^x (1/\log t)\, dt$ |
|---|---|---|
| $10^3$ | 168 | 178 |
| $10^6$ | 78 498 | 78 628 |
| $10^9$ | 50 847 534 | 50 849 235 |

The asymptotic equality (1) was found in Gauss's notes of 1796, without proof, and was never published by Gauss himself. A proof of (1) was first published in 1896 by Hadamard and de la Vallée Poussin. We shall try to make (1) plausible by introducing probabilistic arguments.

## 2. THE SIEVE OF ERATOSTHENES AND THE PRIME NUMBERS

The flowchart given below defines a sieving process, here called the E-process, but known since antiquity under the name of the Sieve of Eratosthenes.

The E-process may be applied to an arbitrary set $M$ of natural numbers and will then produce a subset $M_E$ of $M$; the elements of $M_E$ are called the *prime elements* of $M$.

By applying the E-process to the special set $\{2, 3, 4, \ldots, n\}$, $n$ being a natural number $\geq 2$, the set of the ordinary primes not greater than $n$ is obtained.

### The E-process

[E 1] Let $m$ be the smallest number in the given set $M$ which is not yet cancelled or framed.

[E 2] Frame $m$ (i.e., draw a circle around "$m$")!

[E 3] Cancel all multiples $2m, 3m, 4m, \ldots$ belonging to $M$ (even if these have already been cancelled before).

[E 4] If there are any numbers left in $M$ which are not yet cancelled or framed, go to [E 1].

[E 5] End. The framed numbers are the *Prime Elements* of $M$.

The students should be encouraged to experiment and apply the E-process to different sets $M$.

If the E-process is applied to the special set $M = \{2, 3, 4, \ldots, 100\}$, the situation after four executions of the loop [E 1] – [E 4] is shown by Figure 1.

The next time the loop [E 1] – [E 4] is executed, the number 11 will be framed and its multiples 22, 33, 44, ..., 99 will be cancelled. Whenever during the execution of the E-process a number $m$ is framed, all the numbers preceding $m$ are already either framed or cancelled.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| (2) | (3) | ~~4~~ | (5) | ~~6~~ | (7) | ~~8~~ | ~~9~~ | ~~10~~ | |
| ~~11~~ | ~~12~~ | 13 | ~~14~~ | ~~15~~ | ~~16~~ | 17 | ~~18~~ | 19 | ~~20~~ |
| ~~21~~ | ~~22~~ | 23 | ~~24~~ | ~~25~~ | ~~26~~ | ~~27~~ | ~~28~~ | 29 | ~~30~~ |
| 31 | ~~32~~ | ~~33~~ | ~~34~~ | ~~35~~ | ~~36~~ | 37 | ~~38~~ | ~~39~~ | ~~40~~ |
| 41 | ~~42~~ | 43 | ~~44~~ | ~~45~~ | ~~46~~ | 47 | ~~48~~ | ~~49~~ | ~~50~~ |
| ~~51~~ | ~~52~~ | 53 | ~~54~~ | ~~55~~ | ~~56~~ | ~~57~~ | ~~58~~ | 59 | ~~60~~ |
| 61 | ~~62~~ | ~~63~~ | ~~64~~ | ~~65~~ | ~~66~~ | 67 | ~~68~~ | ~~69~~ | ~~70~~ |
| 71 | ~~72~~ | 73 | ~~74~~ | ~~75~~ | ~~76~~ | ~~77~~ | ~~78~~ | 79 | ~~80~~ |
| ~~81~~ | ~~82~~ | 83 | ~~84~~ | ~~85~~ | ~~86~~ | ~~87~~ | ~~88~~ | 89 | ~~90~~ |
| ~~91~~ | ~~92~~ | ~~93~~ | ~~94~~ | ~~95~~ | ~~96~~ | 97 | ~~98~~ | ~~99~~ | ~~100~~ |

Figure 1.

## 3. A RANDOM SIEVE AND RANDOM PRIMES

The fundamental idea in the subsequent investigations is to replace the E-process of Section 2 by a random sieving process, the SE-process. The E-process can then be considered as a determinate special case of this new random process. The SE-process is defined by the following flowchart and can be applied to an arbitrary set $M$ of natural numbers.

### The SE-process

[SE 1] Let $m$ be the smallest number in the given set $M$ which is not yet cancelled or framed.

[SE 2] Frame $m$ (i.e., draw a circle around "$m$")!

[SE 3] For all $x > m$ and $x \in M$, cancel $x$ *with probability* $1/m$ (even if $x$ has been cancelled before).

[SE 4] If there are any numbers left in $M$ which are not yet cancelled or framed, go to [SE 1].

[SE 5] End. The framed numbers are the *Random Prime Elements* of $M$.

If the SE-process is applied to a set of the form $\{2, 3, 4, \ldots, n\}$, the framed numbers are called random primes. The students should be encouraged to experiment and apply the SE-process to different sets, using tables of random numbers for the random cancellations. A result of the application of the SE-process to the set $\{2, 3, 4, \ldots, 100\}$ is shown by Figure 2.

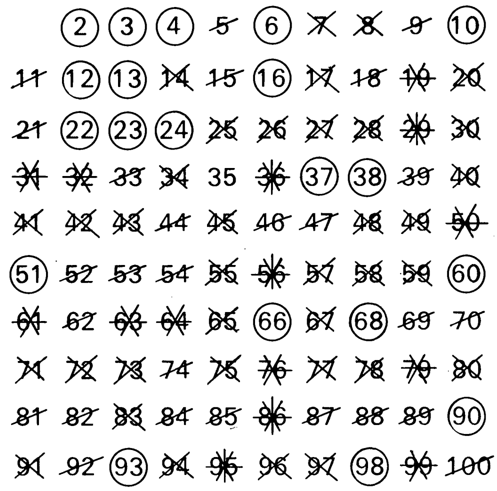

Figure 2.

The determinate E-process and the stochastic SE-process can be considered as *equally effective* sieving processes in the following sense: if a number $m$ happens to be framed during the execution of both processes, then in both cases about $1/m$ of the numbers in any "large" subsequent string of numbers $\{m + 1, m + 2, \ldots, m + n\}$ will be cancelled. By the E-process $[n/m]$ numbers of this string will be cancelled with certainty; by the SE-process numbers will be cancelled at random, but the expected number of cancellations in the string asymptotically equals $n/m$.

## 4. THE EXPECTED NUMBER OF RANDOM PRIMES

Whenever the SE-process is applied to a set of the form $\{2, 3, 4, \ldots, n\}$, ultimately every number in it will be either framed, i.e., be a random prime, or cancelled. Let $X_n$ be the number of random primes thus obtained. $X_n$ is a stochastic variable and its expectation $E(X_n)$, or shortly $E_n$, is the expected number of random primes in the set $\{2, 3, 4, \ldots, n\}$.

Let $p_k$ be the probability that the natural number $k$ is framed, i.e., not cancelled, during an execution of the SE-process. Then $E_n$, the expected number of random primes in {2, 3, 4, ..., $n$}, is given by

$$E_n = \sum_{k=2}^{n} p_k. \qquad (2)$$

Our next step is to derive a nonlinear difference equation by which the probabilities $p_k$ may be determined.

If during an execution of the SE-process the number $k$ happens to be framed, this number $k$ can be conceived as the source of a cancellation wave, called the $k$-wave, which "hits" any subsequent number with probability $1/k$.

In order that a number $m$ be hit by a $k$-wave, two events must take place:

(a) the number $k$ must be framed, which occurs with probability $p_k$;

(b) $m$ must be hit by the cancellation wave emanating from $k$, which occurs with probability $1/k$.

It follows that an arbitrary number $m$ is hit by a $k$-wave, $k < m$, with probability $p_k/k$. The probability that $m$ is not hit by a $k$-wave is then $1 - (p_k/k)$.

Now, the probability that $m$ is framed is $p_m$. But in order to be framed, $m$ must not be hit by any $k$-wave for $k < m$. Therefore

$$p_m = \prod_{k<m}\left(1 - \frac{p_k}{k}\right). \qquad (3)$$

Substituting $m + 1$ for $m$ in (3),

$$p_{m+1} = \prod_{k<m+1}\left(1 - \frac{p_k}{k}\right), \qquad (4)$$

and dividing (4) by (3) yields our fundamental difference equation for the framing probabilities:

$$p_{m+1} = p_m \cdot \left(1 - \frac{p_m}{m}\right) \qquad (5)$$

with initial value $p_2 = 1$. One easily obtains $p_3 = 1/2, p_4 = 5/12, \ldots$. With a computer one can easily compute $E_n$ by using (5) and (2).

## 5. AN ASYMPTOTIC EXPRESSION FOR THE EXPECTED NUMBER OF RANDOM PRIMES

Starting from (2) and (5) one easily arrives at the asymptotic equality

$$E_n \sim \int_2^n (1/\log t)\, dt. \qquad (6)$$

Inverting (5) yields

$$\frac{1}{p_{m+1}} = \frac{m}{p_m(m - p_m)} = \frac{1}{p_m} + \frac{1}{m - p_m}. \qquad (7)$$

Since $0 < p_m < 1$, the following inequality derives from (7):

$$\frac{1}{p_m} + \frac{1}{m} < \frac{1}{p_{m+1}} < \frac{1}{p_m} + \frac{1}{m-1}. \qquad (8)$$

Setting $m = 2, 3, 4, \ldots, n - 1$ in (8) and adding the resulting inequalities yields (observing that $p_2 = 1$):

$$\sum_{m=1}^{n-1}\frac{1}{m} < \frac{1}{p_m} < 1 - \frac{1}{n-1} + \sum_{m=1}^{n-1}\frac{1}{m}. \qquad (9)$$

By using Figure 3 the students readily convince themselves that

$$\sum_{m=1}^{n-1}\frac{1}{m} = \int_2^n \frac{dx}{x} + G_n = \log n + G_n, \qquad (10)$$

where $G_n$ is the area of the shaded domain in Figure 3. The sequence $(G_n)$, obviously being monotonically increasing and satisfying $0 < G_n < 1$ for all $n$, has a limit, and it can be shown that $\lim_{n\to\infty} G_n = \gamma \approx 0.577$, Euler's constant.

Making use of (10), inequality (9) can be rewritten as

$$\log n < \frac{1}{p_n} < 2 + \log n,$$

or equivalently as

$$\frac{1}{2 + \log n} < p_n < \frac{1}{\log n},$$

implying that

$$p_n \sim \frac{1}{\log n}, \qquad (11)$$

i.e., *the probability of n being a random prime is asymptotically* $1/\log n$.

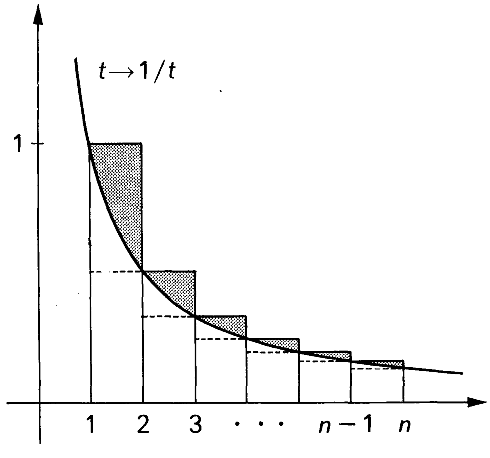


Figure 3.

Making use of (11), the equality (2) can be given the following form:

$$E_n = \sum_{k=2}^{n} p_k \sim \sum_{k=2}^{n} (1/\log k)$$
$$\sim \int_2^n (1/\log t)\, dt \sim n/\log n, \qquad (12)$$

which establishes the Prime Number Theorem for Random Primes.

The observation in Section 3 that the E-process and the SE-process are equally effective makes it plausible (but does not prove) that (12) is valid even for the ordinary primes. Our elementary methods will take us this far.

## 6. THE EXPECTED NUMBER OF DIFFERENT PRIME FACTORS

According to Hardy [4] a number is called "round" if it is the product of a considerable number of comparatively small factors. We here define the *roundness* of a number $n$ as the number of different prime factors of $n$, and try to get an idea of the average roundness of a number.

For this purpose we switch over to an SE-process and random primes. The number of random prime "factors" of a number $n$ equals the number of different $k$-waves hitting $n$. Let $\omega_n$ denote the average number of $k$-waves hitting $n$. Since the probability of $n$ being hit by a $k$-wave is $p_k/k$, the following equality must hold:

$$\omega_n = \sum_{k<n} \frac{p_k}{k}. \qquad (13)$$

Using the results of Section 5,

$$\sum_{k<n} \frac{p_k}{k} \sim \sum_{k<n} \frac{1}{k \cdot \log k}$$
$$\sim \int_2^n \frac{dt}{t \cdot \log t} \sim \text{loglog } n,$$

one arrives at the following asymptotic expression for $\omega_n$:

$$\omega_n \sim \text{loglog } n. \qquad (14)$$

This is the random counterpart of a well-known result by Hardy and Ramanujan of 1917 stating that a composite number $n$ has an average number of loglog $n$ different prime factors.

An excellent survey of the Prime Number Theorem is to be found in [4]. The sources of the heuristic random methods expounded in this article may be found in the literature cited in [3], [2] and [5]. Other elementary and heuristic methods in connection with the Prime Number Theorem are to be found in [1] and [6].

## REFERENCES

1. J. H. Cadwell, *Topics in Recreational Mathematics*, New York: Cambridge University Press, 1970.
2. A. Engel, *Wahrscheinlichkeit und Primzahlen*, *Praxis der Mathematik*, Vol. 5, 1963.
3. P. Erdos, "Some Recent Advances and Current Problems in Number Theory," in T. Saaty, *Lectures in Modern Mathematics*, Vol. 3, New York: John Wiley & Sons, Inc., 1965.
4. Hardy and Wright, *The Theory of Numbers*, Oxford: Oxford University Press, 1945.
5. Mark Kac, *Statistical Independence in Probability, Analysis and Number Theory*, MAA: Carus Mathematical Monograph 12, 1959.
6. Yaglom and Yaglom, *Challenging Mathematical Problems With Elementary Solutions*, San Francisco: Holden-Day, Inc., 1967.