Acta Informatica 17, 477–485 (1982)

Explaining the Wheel Sieve*

Paul Pritchard

Department of Computer Science, Cornell University, Ithaca, New York 14853, USA

Summary. In a previous paper, an algorithm was presented for the classical problem of finding all prime numbers up to a given limit. The algorithm was derived therein by transforming a prior algorithm in accordance with some essentially ad hoc observations on the problem.

The present paper complements the former by developing a simple mathematical framework, which leads to a smoother and more insightful derivation of the new algorithm, and which may be of independent interest to the number theorist.

As if a wheel were in the midst of a wheel.

Ezekiel X:10

I'm just sitting here watching the wheels go round and round; I really love to watch them roll.

John Lennon, Watching the Wheels

1. Introduction

In [8], a solution is presented to the classical problem of finding all prime numbers up to some limit N. The classical solution - the sieve of Eratosthenes - takes $\Theta(N \cdot \log\log N)$ additions; the new solution takes only $\Theta(N/\log\log N)$ additions.

The derivation of the algorithm given in [8] is a distilled history of its discovery, in terms of transformations which improve a prior algorithm. Following Lakatos [4], we believe that such a style of presentation provides useful data for the study of the invention of algorithms[1]. Here, at least, honesty is the best (and easiest) policy.

* A preliminary version of this paper was presented at the fourth Australian Computer Science Conference, Brisbane, 1981

[1] What Lakatos might have called "the logic of computational discovery"

0001-5903/82/0017/0477/$01.80

It is nevertheless important to understand the fundamental nature of our good algorithms. We believe that an algorithm is best explained by first presenting the non-trivial, mathematical theorems needed to prove it correct, then giving an abstract algorithm (with proof), and finally giving the details of its implementation. This paper gives such an explanation of our new prime-number generator.

2. On Wheels

Our algorithm is based on two notions: the sieve and the wheel. The former notion is familiar and plays only a secondary role. To introduce the latter notion, we start with some definitions.

Notation. Variables a, b, m, x etc. range over the positive integers; script variables $\mathcal{S}$, $\mathcal{R}$, $\mathcal{W}$ denote sets of positive integers.

(a,b): the greatest common divisor of a and b.

p_i: the i'th prime number.

Note (e.g. in Fact 5) that the remainder operation **mod** has a lower priority than both addition and multiplication.

Definitions. $\mathcal{R}(m)$: $\{x \mid 1 \leqq x \leqq m \textbf{ and } (x,m)=1\}$
Π_k: $p_1 \cdot p_2 \cdot \ldots \cdot p_k$
$\mathcal{W}_k$: $\mathcal{R}(\Pi_k)$.

$\mathcal{R}(m)$ is known to number theorists as a (particular) reduced residue class $(\text{mod}\, m)$. $\mathcal{W}_k$ is the k'th wheel. Wheels have been in the bag of tricks of the computational number theorist for some time. Two typical used are

(1) in trial-division routines (e.g. in [11], $\mathcal{W}_6$ is "rolled" to generate potential divisors, using the fact that if $x > p_6$, then x can be the smallest prime divisor of a number only if $x \,\mathbf{mod}\, \Pi_6 \in \mathcal{W}_6$), and

(2) in programming Eratosthenes' sieve (e.g. Brent's implementation [1] uses a block of m bytes, where m is a multiple of Π_7, to represent successive sifted intervals of the positive integers - for the k'th interval, the i'th bit of byte n is 0 if and only if the number $(k-1)\cdot m\cdot \text{bytesize}+(i-1)\cdot m+n$ is composite and has been sifted out. Thus if $1 \leqq n \leqq \Pi_7$ and $n \notin \mathcal{W}_7$, all the numbers represented by byte $n+j\cdot\Pi_7$ can be sifted out with a single bit-parallel operation, for $0 \leqq j < m \,\mathbf{div}\, \Pi_7$).

It can be seen from the above uses that a wheel $\mathcal{W}_k$ represents a pattern that repeats modulo Π_k. The result of "rolling a wheel" (indefinitely) is captured by the

Definition. $\mathcal{W}_k^*$: $\{x \mid x \,\mathbf{mod}\, \Pi_k \in \mathcal{W}_k\}$.

We now proceed to establish some simple properties of wheels.

Notation. $a\,..\,b$: $\{x \mid a \leqq x \leqq b\}$.

Primes$(\mathcal{S})$: $\{p \mid p \in \mathcal{S} \textbf{ and } p \text{ is prime}\}$.

next$(\mathscr{S}, a)$: the smallest b such that $b \in \mathscr{S}$ **and** $b > a$
(undefined if no such number exists).

prev$(\mathscr{S}, a)$: the largest b such that $b \in \mathscr{S}$ **and** $b < a$
(undefined if no such number exists).

Fact 1. $1 \in \mathscr{W}_k^*$.

Proof. $(1, \Pi_k) = 1$.

Fact 2. $\text{next}(\mathscr{W}_k^*, 1) = p_{k+1}$.

Proof. p_{k+1} is the smallest $b > 1$ such that $(b, \Pi_k) = 1$.

Fact 3. $\mathscr{W}_k^* \cap p_{k+1} \,..\, p_{k+1}^2 - 1 = \text{Primes}(p_{k+1} \,..\, p_{k+1}^2)$.

Proof. Clearly, $x \in$r.h.s. implies $x \in$l.h.s. Now suppose $x \in$l.h.s. Then x is not divisible by $p_1, \ldots, p_k$ since $x \in \mathscr{W}_k^*$. So the smallest prime divisor p of x satisfies $p \geqq p_{k+1}$. But $x < p_{k+1}^2$, so x must be prime.

Fact 4. $\text{Primes}(1 \,..\, N) = \text{Primes}(1 \,..\, \sqrt{N}) \cup (\mathscr{W}_m^* \cap 1 \,..\, N) - \{1\}$ where p_m is the largest prime $\leqq \sqrt{N}$.

Proof. Immediate from facts 1 to 3, putting $k = m$.

Fact 4 suggests a way to generate Primes$(1 \,..\, N)$, namely by building up $\mathscr{W}_m \cap 1 \,..\, N$ by induction on m. And a standard theorem seems to provide the required step.

Theorem 1. *If* $(m_1, m_2) = 1$, *then*

$$\mathscr{R}(m_1 \cdot m_2) = \{a_2 \cdot m_1 + a_1 \cdot m_2 \, \mathbf{mod} \, m_1 \cdot m_2 \mid a_1 \in \mathscr{R}(m_1), a_2 \in \mathscr{R}(m_2)\}.$$

Proof. In most books on elementary number theory (e.g. [5]).

Fact 5. $\mathscr{W}_{k+1} = \{a \cdot \Pi_k + b \cdot p_{k+1} \, \mathbf{mod} \, \Pi_{k+1} \mid a \in 1 \,..\, p_{k+1} - 1, b \in \mathscr{W}_k\}$.

Proof. Put $m_1 = \Pi_k$, $m_2 = p_{k+1}$ in Theorem 1.

Unfortunately, although Fact 5 shows how $\mathscr{W}_{k+1}$ can be constructed, it does not suggest a way of generating just those (very few, for large k) members of $\mathscr{W}_{k+1}$ that lie in the interval $1 \,..\, N$. An inductive step is instead needed, which can serve to define $\mathscr{W}_{k+1} \cap 1 \,..\, N$ in terms of $\mathscr{W}_k \cap 1 \,..\, N$.

3. The Wheel Sieve

The required characterization can be obtained from the following theorem.

Theorem 2. *Let p be prime. Then*

$$\mathscr{R}(m \cdot p) = \mathscr{R}(m) \cup \{a \cdot m + b \mid a \in \mathscr{R}(p), b \in \mathscr{R}(m)\} - \{p \cdot b \mid b \in \mathscr{R}(m)\}$$

where the set subtraction is needed just when $(p, m) = 1$.

Proof. $\mathscr{R}(m \cdot p) = \{x \mid 1 \leqq x \leqq m \cdot p \, \mathbf{and} \, (x, m \cdot p) = 1\}$

$$= \{x \mid 1 \leqq x \leqq m \cdot p \, \mathbf{and} \, (x, m) = 1\} - \{x \mid 1 \leqq x \leqq m \cdot p \, \mathbf{and} \, (x, p) = p\}$$

$$= \{x \mid 1 \leqq x \leqq m \cdot p \, \mathbf{and} \, (x, m) = 1\} - \{p \cdot b \mid b \in \mathscr{R}(m)\}$$

since a multiple $p \cdot b$ of p need be subtracted if and only if $(b,m)=1$ and $(p,m)=1$

$$=\mathcal{R}(m) \cup \{a \cdot m + b \mid a \in \mathcal{R}(p), b \in \mathcal{R}(m)\} - \{p \cdot b \mid b \in \mathcal{R}(m)\}$$

since the numbers x such that $(x,m)=1$ repeat modulo m.

Fact 6. $\mathcal{W}_{k+1} = \mathcal{W}_k \cup \{a \cdot \Pi_k + b \mid a \in 1 \,..\, p_{k+1}-1, b \in \mathcal{W}_k\} - \{p_{k+1} \cdot b \mid b \in \mathcal{W}_k\}$.

Proof. Put $m = \Pi_k$ and $p = p_{k+1}$ in Theorem 2.

Note that the set union on the r.h.s. of fact 6 is just $\mathcal{W}_k^* \cap 1 \,..\, p_{k+1} \cdot \Pi_k$. Also, fact 6 can be modified simply to define $\mathcal{W}_{k+1} \cap 1 \,..\, N$ in terms of $\mathcal{W}_k \cap 1 \,..\, N$. Facts 2, 4 and 6 immediately suggest the following algorithm, which we dub “the wheel sieve”.

```
$k, p, \mathcal{W}, \Pi, \mathcal{P} := 1, 3, \{1\}, 2, \{2\}$;
  {invariant: $p = p_{k+1}$ and
              $\mathcal{W} = \mathcal{W}_k \cap 1 \,..\, N$ and
              $\Pi = \min\{\Pi_k, N\}$ and
              $\mathcal{P} = \mathrm{Primes}(1 \,..\, p_k)$}
do $p^2 \leqq N \rightarrow$
    Roll $\mathcal{W}$ to $\min\{p \cdot \Pi, N\}$;
    Delete multiples of $p$ from $\mathcal{W}$;
    $\mathcal{P} := \mathcal{P} \cup \{p\}$;
    $k, p := k+1, \mathrm{next}(\mathcal{W}, 1)$
od;
Roll $\mathcal{W}$ to $N$
{$\mathcal{W} \cup \mathcal{P} - \{1\} = \mathrm{Primes}(1 \,..\, N)$}
```

The procedure “Roll $\mathcal{W}$ to n” must satisfy

$$\{\mathcal{W} = \mathcal{W}_k \cap 1 \,..\, N \textbf{ and } \Pi = \min\{\Pi_k, N\} \textbf{ and } N \geqq n \geqq \Pi\}$$

$$\text{Roll } \mathcal{W} \text{ to } n \ \{\mathcal{W} = \mathcal{W}_k^* \cap 1 \,..\, n \textbf{ and } \Pi = n\},$$

and may only change $\mathcal{W}$ and Π. Refining the procedure is straightforward (though notice that $x = a \cdot \Pi + b$ is computed with $x := \Pi + \beta$, where $\beta = (a-1) \cdot \Pi + b$ has previously been added):

```
$x, \beta := \Pi + 1, 1$;
do $x \leqq n \rightarrow \mathcal{W} := \mathcal{W} \cup \{x\}$;
                $\beta := \mathrm{next}(\mathcal{W}, \beta)$;
                $x := \Pi + \beta$
od;
$\Pi := n$.
```

The multiples of p that must be deleted from $\mathcal{W}$ are, by Fact 6, those in the set $\{p \cdot b \mid b \in \mathcal{W}\}$. These can be deleted one at a time provided that a multiple $p \cdot b$ never fails to be deleted because the factor b is a previously deleted multiple of p. The solution is simple – the factors b are taken in decreasing order. The first such factor can be obtained by a preliminary forward search.

```
$b:=p$;
do $p\cdot b \leqq \Pi \to b:=\mathrm{next}(\mathcal{W},b)$ od;
do $b>1 \to b:=\mathrm{prev}(\mathcal{W},b)$;
        $\mathcal{W}:=\mathcal{W}-\{p\cdot b\}$
od
```

Proving the correctness of the wheel sieve above is an easy exercise given Facts 2, 4 and 6, provided that all instances of operations next and prev are defined. (This needs checking since $\mathcal{W}=\mathcal{W}_k\cap 1\,..\,N$, not $\mathcal{W}=\mathcal{W}_k^*$, is invariant.) So consider firstly the refinement of "Roll $\mathcal{W}$ to n". Operation next$(\mathcal{W},\beta)$ is defined since the previous assignment adds x to $\mathcal{W}$ and $x>\beta$. Next consider the code for deleting multiples of p – next$(\mathcal{W},b)$ is defined since the truth of the guard ensures that $p\cdot b\in\mathcal{W}$; similarly prev$(\mathcal{W},b)$ is defined since $1\in\mathcal{W}$. The last operation to check is next$(\mathcal{W},1)$ in the top level of the algorithm. There are two cases to consider. First, if $\Pi=\Pi_{k+1}$, then $\Pi-1\in\mathcal{W}$ and, since $k+1>1$, $\Pi-1>1$. Otherwise, $\Pi=N<\Pi_{k+1}$ and we can appeal to a theorem that guarantees that, for all i, a prime exists between p_i and p_i^2. Since $p^2\leqq N$, this guarantees that $p_{k+2}\in\mathcal{W}$ as required. (But note that it is easy to modify the algorithm so that its correctness no longer depends on this deeper result – it is sufficient to maintain a "sentinel", such as $N+1$, in $\mathcal{W}$.)

The precondition for the above algorithm may thus be safely taken as $N\geqq 2$. It is interesting to note that the version derived in a more intuitive, less mathematical fashion in [8] fails for $N=3,4$ since the operation next $(\mathcal{W},1)$ becomes undefined. The reader is referred to [8] for implementations leading to a complexity of $\Theta(N/\log\log N)$ additions.

4. Wheels Within Wheels

There is an amusing and instructive geometric/pictorial model of Fact 6 "in action". The idea is to draw wheel $\mathcal{W}_k$ as a circle C_k of perimeter Π_k, representing each member x of $\mathcal{W}_k$ with a mark x units clockwise along the perimeter from an origin. The wheels share a common centre, and their origins lie on a vertical line on the same side of the centre.

C_{k+1} is obtained from C_k in a 2-phase operation. First, an unmarked circle C is drawn for $\mathcal{W}_{k+1}$. This should have p times the perimeter of C_k, where p is the first marked point of C_k after 1. Then the inner circle C_k in moved up until its origin coincides with that of C. C_k is then rolled clockwise around the inner perimeter of C, and a mark is left wherever a marked point of C_k touches C. This first phase corresponds to rolling $\mathcal{W}_k$, or, more formally, to the set union on the r.h.s. of Fact 6.

The second phase deletes the multiples of p. For this, the inner circle C_k is returned to its usual position, and lines are drawn radially outward from each marked point on C_k. These touch the outer circle at exactly those points whose marks should be erased to leave the circle C_{k+1}!

Providing "the first marked point of C_k after 1" is taken to cover the cases (for $k=0,1$) where it is necessary to go right round the circle, this model of wheel building can start with a circle C_0 of length 1 with a mark at the origin!

Figure 1 below shows the construction of C_3 from C_2.

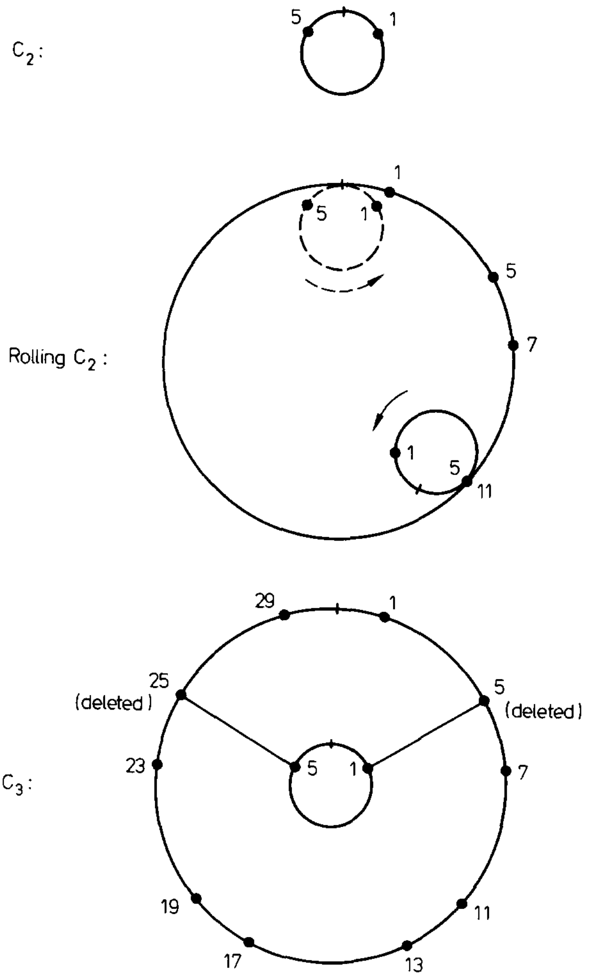

Fig. 1. Constructing C_3 from C_2

5. Drawing a Theorem

In the transition from $\mathscr{W}_k$ to $\mathscr{W}_{k+1}$, the set of multiples deleted after rolling $\mathscr{W}_k$ is $\{p_{k+1}\cdot b \mid b\in\mathscr{W}_k\}$. It is a standard result of elementary number theory (see [5]) that if $(m_1, m_2)=1$ then $\{m_2\cdot b \mathbin{\mathbf{mod}} m_1 \mid b\in\mathscr{R}(m_1)\}=\mathscr{R}(m_1)$, so exactly one representative $(\operatorname{mod} \Pi_k)$ of each member b of $\mathscr{W}_k$ is deleted.

Let g_k be the maximum gap between successive elements of $\mathscr{W}_k^*$ (ordered numerically). We give a simple diagrammatic proof of the following (new?) result.

Theorem 3. $g_k \geqq 2p_{k-1}$, for all $k>1$.

Proof. For $k>2$, C_k has the form

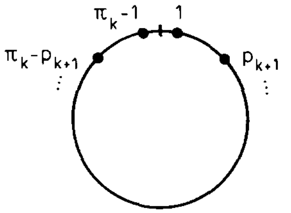

Fig. 2

So some part of C_{k+1} has (by construction) the form

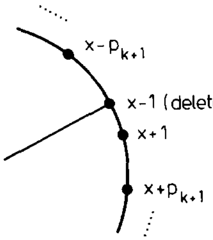

Fig. 3

Continuing, some part of C_{k+2} has the form

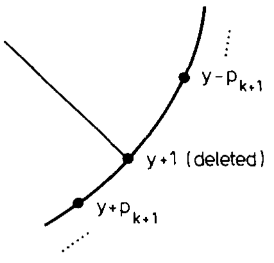

Fig. 4

That is, $g_{k+2} \geqq 2p_{k+1}$ for all $k>2$.
An easy computation of g_2, g_3 and g_4 completes the proof.

As an application of Theorem 3, we answer a question of Kanold [3]. Define $g(m)$ as the least integer such that every set of $g(m)$ consecutive integers contains a number x such that $(x,m)=1$. Kanold asked whether, if m is the product of the primes less than d that do not divide d, it is true that $g(m)<d$.

Let us put $d=p_{k+1}$, so that $m=\Pi_k$ and $g(m)=g_k \geqq 2p_{k-1}$ by Theorem 3. But it is known (e.g. see [10]) that $p_{i+2}<2p_i$ for $i>3$. So for $k>4$, $g_k>p_{k+1}$, and Kanold's question is answered in the negative. This is a considerably simpler answer than that given by Pomerance [7].

6. On Misra's explanation of a linear sieve

Misra [6] explains a linear sieve algorithm that is based on the invariant[2]

$$p=p_{k+1} \quad \textbf{and} \quad \mathcal{W}=\mathcal{W}_k^* \cap 1\,..\,N \cup \mathrm{Primes}(1\,..\,p).$$

His method of deleting multiples differs from ours in that a forward search for the largest multiple $p \cdot r$ to be deleted is avoided by keeping invariant the relation

$$r=\text{the largest number in } \mathcal{W} \text{ such that } p \cdot r \leqq N.$$

To show these invariants imply that $\mathrm{next}(\mathcal{W}, r)$ is defined, an appeal is made to the "rather deep" theorem of Chebyshev, that for any $i>1$ there is a prime q with $i<q<2i$. The appeal is unnecessary. For if $r<p$, then $\mathrm{next}(\mathcal{W}, r)$ is defined since $p \in \mathcal{W}$; otherwise, $r \geqq p$, so $r \in \mathcal{W}_k^*$ and $\mathrm{next}(\mathcal{W}, r)$ is defined since $p \cdot r \in \mathcal{W}$.

However, as in our algorithm, it is necessary to know that $p_{i+1}<p_i^2$ for all i to show that $\mathrm{next}(\mathcal{W}, p)$ is defined after the deletions have been done. This is not stated in [6].

It is conjectured in [6] that r and $\mathrm{prev}(\mathcal{W}, r)$ cannot both be multiples of p, and it is shown how advantage may be taken if this is true. The conjecture is equivalent to the assertion that no gap on C_k is a multiple of p_{k+1}. Since any gap on C_k is carried over to C_{k+1}, the conjecture is falsified if any gap on C_k is a multiple of p_{k+i} for some $i \geqq 1$. We have shown that g_k, the maximum gap on C_k, is at least $2p_{k-1}$. As noted in [2], a result of Rankin gives a constant c such that

$$g_k \geqq c \cdot \log^2 k \cdot \mathrm{logloglog}\, k \cdot (\mathrm{loglog}\, k)^{-2} \qquad \text{for } k>e^{e^e}.$$

But since $p_i \sim i \cdot \log i$, it follows that $p_{k+1}=o(g_k)$. Given these facts, we think it reasonable to conjecture that Misra's conjecture is false.

7. Conclusion

A good case, we would argue, has been made for the computational and mathematical utility of wheels. Furthermore our proof of Theorem 3 is a nice example of how the algorithmic considerations of the computer scientist can produce insight of genuine mathematical interest (see [9] for another example).

Finally, in case the reader has observed that wheels are symmetric and that use might be made of this in our algorithm, we point out that we have had no

[2] We write $\mathcal{W}$, N for the S, n of [6]

truck with semi-wheels on the grounds of elegance and efficiency. A semi-wheel sieve is more complex, and no more efficient asymptotically, than the wheel sieve.

References

1. Brent, R.P.: The first occurrence of large gaps between successive primes. Maths. of Comp. **27**, 959-963 (1973)
2. Erdös, P., Penney, D.E., Pomerance, C.: On a class of relatively prime sequences. J. Number Theory **10**, 451-474 (1978)
3. Kanold, H.J.: Über Primzahlen in arithmetischen Folgen. Math. Ann. **156**, 393-395 (1964); II, Math. Ann. **157**, 358-362 (1965)
4. Lakatos, I.: Proofs and refutations. Cambridge: Cambridge University Press 1976
5. Le Veque, W.J.: Elementary theory of numbers. Reading, Ma: Addison-Wesley 1965
6. Misra, J.: An exercise in program explanation. ACM Trans. Program. Lang. Syst. **3**, 104-109 (1981)
7. Pomerance, C.: A note on the least prime in an arithmetic progression. J. Number Theory **12**, 218-223 (1980)
8. Pritchard, P.: A sublinear additive sieve for finding prime numbers. Comm. ACM **24**, 18-23 (1981)
9. Pritchard, P.: Another look at the "longest ascending subsequence" problem. Acta Informat. **16**, 87-91 (1981)
10. Sierpiński, W.: Elementary theory of numbers. Warsaw: Pańtwowe Wydawnictwo Naukowe 1964
11. Wunderlick, M.D., Selfridge, J.L.: A design for a number theory package with an optimized trial division routine. Commun. ACM **17**, 272-277 (1974)

Received July 27, 1981/July 16, 1982