Cluster Primes

Richard Blecksmith, Paul Erdős, and J. L. Selfridge

1. INTRODUCTION. A prime $p > 2$ is called a *cluster prime* if every even positive integer less than $p - 2$ can be written as a difference of two primes $q - q'$, where q and q' are both less than or equal to p. Due to the concentration of primes at the beginning of the positive numbers, the first 23 odd primes 3, 5, 7, $11, \ldots, 89$ are all cluster primes. The smallest non-cluster prime is 97: the previous prime is 89 and so $88 = 97 - 9$ is not a difference of two primes smaller than 98. In general, if p is a cluster prime, then there must be enough primes in a "small" neighborhood to the left of p so that the even numbers $p - 9$, $p - 15$, $p - 21$, $p - 25$, etc., can all be written as the difference of primes less than p.

The notion of cluster primes is reminiscent of the notion of prime constellations. The most famous prime constellation is that of the twin primes $\{p - 2, p\}$. More elaborate constellations such as $\{p - 8, p - 6, p - 2, p\}$ have also been studied [**4**, pp. 64–68]. In a prime constellation both the number of primes and the differences between them are fixed. There is no guarantee that the largest prime in a prime constellation is a cluster prime. A sparse set of primes may lie just in front of this constellation. The first pair of twin primes $\{p - 2, p\}$ for which p is *not* a cluster prime is $\{227, 229\}$, because the number $p - 27 = 202$ is not a difference of primes less than 230. To see this, observe that $202 = 211 - 9 = 223 - 21 = 227 - 25$, and 211, 223, 227 (the twin of 229) are the only primes between 202 and 228. We show in the proof of Theorem 1 that the number of primes in a small interval just before a cluster prime p grows in size with p. In this way, a cluster prime can be thought of as the largest prime in a "galaxy" of primes.

It is reasonable to expect that among the primes, the cluster primes become increasingly rare. In spite of the initial head start of 23 consecutive cluster primes, the non-cluster primes quickly catch up, so that by the prime 2251 we have 167 cluster primes and 167 non-cluster primes. Starting with 2267, the next prime after 2251, the cluster primes begin to lag further and further behind. When we reach 10^{13}, the non-cluster primes outnumber the cluster primes by a ratio of about 325 to 1.

The simplest question we can ask about the distribution of cluster primes is:

Are there infinitely many cluster primes?

An affirmative answer would imply that $p_{n+1} - p_n \leq 6$ for infinitely many primes p_n, which is a well-known hopeless problem. We enjoy more success looking for an *upper* bound for $\pi_c(x)$, the number of cluster primes not exceeding x. In Section 2 we show that eventually $\pi_c(x)$ is less than $x/(\log x)^s$, for any fixed positive integer s. Our result "suggests" that the cluster primes are less numerous than the twin primes, although we have no way of proving that either of these two collections is infinite. Our theorem is powerful enough, however, to show that the sum of the reciprocals of the cluster primes converges, a result well-known for the twin primes.

In Section 3 we present an efficient algorithm, which, given a particular cluster prime p_n, determines the next cluster prime greater than p_n. We used this algorithm to compute the cluster primes up to 10^{13} and we give the values of $\pi_c(x)$ for powers $x = 10^k$, where k ranges from 2 to 13. These data and a discussion of the results are presented in Section 4. We conclude with a comparison of the number of cluster primes versus the number of pairs of twin primes.

2. AN UPPER BOUND FOR $\pi_c(x)$. We have the following result:

Theorem 1. *For every positive integer s, there is a bound $x_0 = x_0(s)$ such that if $x \geq x_0$ then*

$$\pi_c(x) < \frac{x}{(\log x)^s}. \tag{1}$$

The proof is based on the following two lemmas:

Lemma 1. *Let $\pi(x)$ denote the number of primes $\leq x$. Then for $x \geq 6$,*

$$\pi(x) < \frac{2x-6}{\log x}.$$

Proof: We use the estimate of Rosser and Schoenfeld [**5**]:

$$\pi(x) < \frac{1.256x}{\log x}, \qquad x > 1. \tag{2}$$

Since

$$\frac{2x-6}{\log x} > \frac{1.256x + (0.7x - 6)}{\log x} > \frac{1.256x}{\log x}$$

for all $x \geq 9$, Lemma 1 follows for $x \geq 9$. One can easily verify that the formula in the lemma holds for $6 \leq x < 9$. ∎

Our main tool in proving Theorem 1 is the following application of Brun's sieve. The notation $f(x) \ll g(x)$ means that for some constant M and value x_0, $|f(x)| \leq Mg(x)$ for all $x \geq x_0$.

Lemma 2. *Let s be a natural number, let $d_1, \ldots, d_s$ be s distinct, nonzero integers, and let $f(x)$ count the number of primes p in the interval $0 < p \leq x$ such that the differences $p - d_i$ are prime for each $i = 1, \ldots, s$. Then*

$$f(x) \ll \prod_{p \mid \prod_{1 \leq i < j \leq s}(d_i - d_j)} \left(1 - \frac{1}{p}\right)^{\rho(p)-s} \prod_{p \mid d_1 \cdots d_s} \left(1 - \frac{1}{p}\right)^{-1} \frac{x}{(\log x)^{s+1}},$$

where $\rho(p)$ denotes the number of modulo p distinct numbers among the d_i's, and where the constant implied by the $\ll$-notation depends only on s.

For a proof of this lemma, take $y = x$ in Corollary 2.4.2 in [**2**, p. 81].

Proof of Theorem 1: Suppose p is a cluster prime. We wish to get a lower bound on the number of primes in the interval $[p - t, p)$, where t is a small positive integer, to be specified later. By the definition of cluster prime, every even number $2r$ in the interval $p - t \leq 2r \leq p - 3$ must be of the form $q - q'$, where q and q' are primes $\leq p$. Clearly q' must be $\leq t$. By Lemma 1, if $t \geq 6$ the number of these primes q' is less than $2(t-3)/\log t$. On the other hand, there are more than

$(t-3)/2$ even numbers in the interval $[p-t, p-3]$. Thus there must be at least $\frac{1}{4}\log t$ primes in $[p-t, p)$. Define

$$s = \left\lfloor \frac{\log t}{4} \right\rfloor. \tag{3}$$

There are $\binom{t}{s}$ ways to place s numbers $q_1 > q_2 > \cdots > q_s$ in the interval $[p-t, p)$. (We allow the q_i to be even to simplify the calculations.) Using the crude estimate $\binom{t}{s} \le nt^s$, there are fewer than t^s choices for the s differences $d_i = p - q_i$, $1 \le i \le s$. If the differences $d_1 < \cdots < d_s$ are fixed, Lemma 2 ensures that the number of choices of $p \le x$ such that each $p - d_i$ is prime is at most $Mx/(\log x)^{s+1}$, where M is a constant depending only on s. Thus

$$\pi_c(x) < M\frac{x}{(\log x)^{s+1}}t^s.$$

Now given s, let t be the least positive integer satisfying equation (3). Taking x so large that $t^s \le \log x$, we have

$$\pi_c(x) < M\frac{x}{(\log x)^s}.$$

Theorem 1 follows easily. ■

A consequence of Theorem 1 is the following result.

Theorem 2. *The sum of the reciprocals of the cluster primes is finite.*

Proof: If the set of cluster primes is finite, there is nothing to prove, so assume there are infinitely many, and denote the n-th cluster prime by q_n. Consider Theorem 1 with $s = 2$. For n sufficiently large,

$$\pi_c(q_n) < \frac{q_n}{(\log q_n)^2}.$$

But $\pi_c(q_n) = n$ and $(\log q_n)^2 > (\log n)^2$. Thus, for n sufficiently large, we have $q_n > n(\log n)^2$. Since the series $\Sigma n^{-1}(\log n)^{-2}$ converges, by the integral test, it follows that $\Sigma 1/q_n$ converges by the comparison test. ■

It appears that a stronger result than Theorem 1 may actually be true:

Conjecture. *For some constant α, we have*

$$\pi_c(x) \ll \frac{x}{e^{\alpha(\log\log x)^2}}. \tag{4}$$

This result would follow from Lemma 2 if we could guarantee that the implied constant in the lemma does not grow too fast as a function of s.

3. THE ALGORITHM. Let p_n be the nth prime. We describe an algorithm that inputs the index n of the current cluster prime p_n and returns the index of the next cluster prime. The idea is simple. Since p_n is a cluster prime, we know that every even integer from 2 to $p_n - 3$ can be expressed as a difference of two primes not greater than p_n. In order to check whether the next prime p_{n+1} is also a cluster prime, we need examine only the even numbers $p_n - 1$ through $p_{n+1} - 3$. If $p_{n+1} - p_n = 2$, 4, or 6, then there is nothing to check; p_{n+1} is the next cluster prime. If $p_{n+1} - p_n \ge 8$, then p_{n+1} is a non-cluster prime, since $p_{n+1} - 9$ cannot be a difference of two smaller primes. In this case we examine the even numbers

$p_{n+1} - t$, where t is an odd composite less than or equal to $p_{n+1} - p_n + 1$. For each such t we look ahead in the sequence of primes $\{p_{n+m+1}\}_{m=1}^{\infty}$ until $q' = t + p_{n+m+1} - p_{n+1}$ turns out to be prime. Here lies the significance of little m (as opposed to capital M) in the algorithm: p_{n+m+1} is the *first* prime for which the *particular* even number $p_{n+1} - t$ can be written as a difference $p_{n+1+m} - q'$ of primes; so we are at least m primes away from the next cluster prime at this stage of the algorithm. Capital M is the maximum number of primes to the next (possible) cluster prime, based on all of the previous values of m found so far. When we move on to the *next* prime at the end of the outer do-loop, we decrease M by 1, since we are now one prime closer to the cluster prime we are seeking. We continue processing consecutive primes until $M = 0$, indicating that we have finally reached the next cluster prime. For example, the next prime after the non-cluster prime $p_{25} = 97$ is $p_{26} = 101$. Since $p_{25} - 9 = p_{26} - 13$, 101 is the next cluster prime after 89. In our algorithm, we need a list of the prime differences $\text{diff}[n] = p_{n+1} - p_n$. We do not require the actual values of the primes themselves, just the differences, since

$$p_{n+m+1} - p_{n+1} = \sum_{i=1}^{m} \text{diff}[n+i].$$

We also need a short list, named odd_comp, of the odd composites 9, 15, 21, 25, 27, 33, etc., as well as a look-up table to tell when a "small" odd integer is prime.

Algorithm Find_next_cluster__prime(n, current_prime).

```
M = 0;
do
  d = diff[n];
  if (d > 6)
    for (i = 1; odd_comp[i] ≤ d + 1; i = i + 1)
      m = 0;
      t = odd_comp[i];
      repeat
        m = m + 1;
        t = t + diff[n + m];
      until t is prime;
      if (m > M) M = m;
  n = n + 1;
  current_prime = current_prime + d;
  M = M − 1;
while (M > 0)
return n;
```

It is worth pointing out that the efficiency of this algorithm is due to the fact that it always looks forward and *never needs to backtrack*. In actual practice, the program spends more time sieving for the prime differences than it does running the algorithm.

4. DISTRIBUTION OF CLUSTER PRIMES UP TO 10^{13}. We encoded this algorithm in a C program and ran it on a 300 MHz Sun Ultra 2 Workstation. Our goal was to tabulate the cluster primes up to 10^{13} in order to get an indication of their distribution. The following short table gives the number of cluster primes versus non-cluster primes for powers of 10. Here $\pi_c(x)$, $\pi_n(x)$, and $\pi_2(x)$, respectively, denote the number of cluster primes, non-cluster primes, and twin prime constellations less than or equal to x. Since we do not count 2 as a cluster or non-cluster prime, we have the equation $\pi_c(x) + \pi_n(x) + 1 = \pi(x)$, which we can use as a check on the data. The values of $\pi_2(x)$ were computed by Brent [**1**] and can also be found in [**2**, p. 262]. The last column gives the value of α for which (4) becomes an equality, viz. $\alpha = \log(x/\pi_c(x))/(\log\log x)^2$.

x	$\pi_c(x)$	$\pi_n(x)$	$\frac{\pi_n(x)}{\pi_c(x)}$	$\pi_2(x)$	α
10^2	23	1	0.04	8	0.6301
10^3	99	68	0.69	35	0.6192
10^4	420	808	1.92	205	0.6430
10^5	1807	7784	4.31	1224	0.6722
10^6	8287	70,210	8.47	8169	0.6952
10^7	40,017	624,561	15.61	58,980	0.7144
10^8	202,208	5,559,246	27.49	440,312	0.7308
10^9	1,059,812	49,787,721	46.98	3,424,506	0.7455
10^{10}	5,736,857	449,315,654	78.32	27,412,679	0.7586
10^{11}	31,914,282	4,086,140,530	128.03	224,376,048	0.7707
10^{12}	182,065,897	37,425,846,120	205.56	1,870,585,220	0.7817
10^{13}	1,061,375,739	345,004,161,099	325.05	15,834,664,872	0.7921

By $x = 10^4$ the non-cluster primes outnumber the cluster primes by a ratio of roughly 2 to 1. As expected, the ratio $\pi_n(10^k)/\pi_c(10^k)$ increases for each exponent k, and when we reach 10^{13}, approximately 0.3% of the primes are cluster primes. We can show this behavior by using Theorem 1 together with the prime number theorem, which states that $\pi(x)$, the number of primes not greater than x, is asymptotic to $x/\log x$. Since $\pi_c(x)$ is eventually less than $x/(\log x)^2$, it follows that the ratio $\pi_c(x)/\pi(x)$ approaches 0 as x tends to infinity. Thus, "most" primes are non-cluster primes and the ratio $\pi_n(x)/\pi_c(x) = (\pi(x) - 1)(\pi_c(x))^{-1} - 1$ must tend to infinity.

It is interesting to contrast the columns for π_n and π_c. The ratios $\pi_n(10^{k+1})/\pi_n(10^k)$ seem to be approaching the limit 10. A proof of this observation follows from the fact that $\pi_n(x)$ is asymptotic to $\pi(x)$ and from the prime number theorem:

$$\lim_{k\to\infty} \frac{\pi_n(10^{k+1})}{\pi_n(10^k)} = \lim_{k\to\infty} \frac{\pi(10^{k+1})}{\pi(10^k)} = \lim_{k\to\infty} \frac{10^{k+1}}{(k+1)\log 10} \Big/ \frac{10^k}{k \log 10} = 10.$$

For the cluster primes, the ratios $\pi_c(10^{k+1})/\pi_c(10^k)$ increase from 4.24 for $k = 3$ to 5.83 for $k = 12$. It is difficult to predict a limit from such limited data. Heuristic considerations, however, suggest that $\pi_c(x)$ has the shape $x^{1-h(x)}$, where $h(x)$ is a function whose limit tends to 0 as x goes to infinity. If this estimate is correct, then the ratios $\pi_c(10^{k+1})/\pi_c(10^k)$ would also tend to 10, though more slowly than the ratios for the non-cluster primes.

As the program ran to 10^{13}, the largest value of M in the algorithm Find_next_cluster_prime was 58 and the largest value of t was 1503. The largest number of consecutive non-cluster primes was 10,543, found between the cluster primes 8,353,771,390,333 and 8,353,771,707,107. The difference between these two primes is 316,774, the largest difference found up to 10^{13}.

It is worthwhile noting that past 10^6, the number of cluster primes lags behind the number of twin primes. For $x = 10^{12}$ the number of twin primes is roughly ten times larger than the number of cluster primes. To explain this phenomenon, put

$s = 1$ and $d_1 = 2$ into Lemma 2 (Brun's sieve) to get the upper bound

$$\pi_2(x) \ll \frac{x}{(\log x)^2}. \tag{5}$$

Brun used this estimate in 1921 to show that the sum of the reciprocals of the twin primes converges; the proof is essentially the same as our proof of Theorem 2. Comparing (5) with Theorem 1's estimate $\pi_c(x) \ll x/(\log x)^s$ for any positive integer s, we would expect the cluster primes to be rarer than the twin primes. In the interest of honesty, however, we must admit two facts. First, the estimate in (2) for $\pi_c(x)$ holds for $x \geq x_o(s)$. On examining the proof of Theorem 1, the value of $x_0(s)$ is roughly $x_0(s) = e^{t^s}$, where $t = e^{4s}$. For $s = 3$, this bound is $x_0 = e^{e^{36}}$, a number with approximately 1.87×10^{15} decimal digits. It seems a bit presumptuous to think that we are seeing the effects of Theorem 1 with $s = 3$ for the comparatively small 13 digit numbers. The second remark is that although upper bounds may indicate what happens, they are not conclusive. For all we know, the number of cluster primes and twin primes could both be finite. In 1922 Hardy and Littlewood conjectured that $\pi_2(x)$ is asymptotic to

$$2\prod_{p\geq 3}\frac{p(p-2)}{(p-1)^2}\int_2^x\frac{dx}{(\log x)^2} \approx 1.320323632\int_2^x\frac{dx}{(\log x)^2}.$$

This famous conjecture has been shown to be remarkably accurate in estimating the number of twin prime constellations [**3**, p. 66], and strengthens our belief that $\pi_2(x) arrow \infty$ with x. That $\pi_c(x)$ tends to infinity seems harder to prove.

ACKNOWLEDGMENTS. We thank Heini Halberstam for helping elucidate the phrase "by Brun's sieve" in Paul Erdős' handwritten proof of Theorem 1 in Section 2. We are also grateful for discussions with David Rusin. Michael Filaseta is responsible for many improvements throughout the manuscript. Finally, Eric Behr helped with the systems requirements for implementing the cluster prime algorithm.

REFERENCES

1. R. P. Brent, Irregularities in the distribution of primes and twin primes, *Math. Comp.* **29** (1975) 43–56.
2. H. Halberstam and H. E. Richert, *Sieve Methods*, Academic Press, 1974.
3. Paulo Ribenboim, *The New Book of Prime Number Records*, Springer, 1996.
4. Hans Riesel, *Prime Numbers and Computer Methods for Factorization*, Birkhäuser, 1985.
5. J. Barkley Rosser and L. Schoenfeld, Approximate formulas for some functions of prime numbers, *Illinois J. Math.* **6** (1962) 64–97.

RICHARD BLECKSMITH grew up in Cincinnati, Ohio and has been a longtime fan of baseball (Reds), basketball (Bulls), motorcycles (Harleys), and mathematics (Number Theory). He earned his PhD from the University of Arizona in 1983 under the direction of John Brillhart and is a professor at Northern Illinois University. With his wife Sharon, daughter Katie and various itinerate mathematicians, he lives on a small farm near Malta, Illinois, where he is often out standing in his field. He is Secretary of the Number Theory Foundation and is also Vice President of the Malta Band Boosters.
Department of Mathematical Sciences, Northern Illinois University, DeKalb, Illinois 60115
richard@math.niu.edu

J. L. SELFRIDGE was born and raised in Ketchikan, Alaska. He earned his BS in Math at the University of Washington and his PhD at U.C.L.A. He found factors of the Fermat numbers F_{16} and F_{10} in 1953 and has solved various Erdős problems for 40 years, culminating in about 15 joint papers. Pomerance, Selfridge, and Wagstaff offer $620 for a counter-example or proof of an easy primality test. (See page 28 of Richard Guy's *Unsolved Problems in Number Theory*, 2nd edition.) His life included 15 exciting years of full-time administration. Retirement since 1991 has been a mathematical feast.
Department of Mathematical Sciences, Northern Illinois University, DeKalb, Illinois 60115
selfridg@math.niu.edu