MATHEMATICS OF COMPUTATION
VOLUME 61, NUMBER 203
JULY 1993, PAGES 69–82

STATISTICAL EVIDENCE FOR SMALL GENERATING SETS

ERIC BACH AND LORENZ HUELSBERGEN

Dedicated to the memory of D. H. Lehmer

ABSTRACT. For an integer n, let $G(n)$ denote the smallest x such that the primes $\leq x$ generate the multiplicative group modulo n. We offer heuristic arguments and numerical data supporting the idea that

$$G(n) \leq (\log 2)^{-1} \log n \log\log n$$

asymptotically. We believe that the coefficient $1/\log 2$ is optimal. Finally, we show the average value of $G(n)$ for $n \leq N$ is at least

$$(1 + o(1)) \log\log N \log\log\log N,$$

and give a heuristic argument that this is also an upper bound. This work gives additional evidence, independent of the ERH, that primality testing can be done in deterministic polynomial time; if our bound on $G(n)$ is correct, there is a deterministic primality test using $O(\log n)^2$ multiplications modulo n.

1. INTRODUCTION

The purpose of this paper is to study the behavior of

$$G(n) = \min\{x : \mathbf{Z}_n^* \text{ is generated by primes } \leq x\}.$$

(Here, $\mathbf{Z}_n^*$ denotes the multiplicative group of residue classes mod n.) We give heuristic arguments to suggest that

$$\limsup_{n\to\infty} \frac{G(n)}{\log n \log\log n} = \frac{1}{\log 2},$$

and present empirical results in support of the above estimate.

Our interest in the asymptotic behavior of $G(n)$ stems from the analysis of several number-theoretic algorithms. For example, to test an integer n for primality, it suffices to use the strong pseudoprime test [26, 29] with prime bases up to $G(n)$; if the above conjecture is true, then there is a deterministic primality test using $O(\log n)^2$ multiplications modulo n. Another example is provided by Tonelli's algorithm [33] for computing square roots mod p. This algorithm uses a number a satisfying $(a|p) = -1$; because this condition is easy to test when a is small, it is common to simply use an $a \leq G(n)$. This is done, for example, in the computer algebra system *Mathematica*.

Our heuristic model can be summarized as follows: we assume that the small primes $p = 2, 3, 5, \ldots$ represent independent samples from $\mathbf{Z}_n^*$. Probability

Received by the editor July 10, 1992 and, in revised form, November 2, 1992.
1991 *Mathematics Subject Classification.* Primary 11-04; Secondary 11N25, 11Y11, 11Y70.

©1993 American Mathematical Society
0025-5718/93 $1.00 + $.25 per page

theory implies that the number of primes needed to generate $\mathbf{Z}_n^*$ is almost surely less than $(1 + o(1)) \log_2 n$, and our main conjecture comes from applying the prime number theorem to this bound. A matching lower bound arises from a different heuristic argument. To test this, we computed $G(n)$ for every $n \leq 10^9$; our data suggest that the theory is correct. For example, there are 23 values of n in this range for which $G(n)$ exceeds any previous value; for these n, the ratio of $G(n)$ to the $\lfloor \log_2 n \rfloor$th prime lies between 1 and 1.86, with an average value of 1.187.

Before proceeding further, we review what is actually known about $G(n)$. The Pólya-Vinogradov inequality [8], combined with a lower bound for the number of small units $\bmod n$ [34], implies that $G(n) = O(\sqrt{n} \log n \log\log n)$. There does not seem to be a better general bound, although this can be sharpened in special cases [21]. In particular, for prime n we have $G(n) = n^{1/4+o(1)}$ [5]. If the ERH is true, then $G(n) = O(\log n)^2$ [28].

We can obtain lower bounds on the growth rate of $G(n)$ by observing that $G(n)$ is, for odd n, at least as large as the least q with $(q|n) = -1$. Graham and Ringrose [15] proved that the least quadratic nonresidue modulo a prime p is $\Omega(\log p \log\log\log p)$, improving the Friedlender-Salié bound of $\Omega(\log p)$. (Here we use the Ω-notation in the sense of Hardy, to indicate that a constant times $\log p \log\log\log p$ is exceeded infinitely often.) Montgomery [28] showed that if the ERH is true, this can be raised to $\Omega(\log p \log\log p)$. Thus, $G(n)$ is occasionally a bit larger than $\log n$, and presumably no more than a constant times $(\log n)^2$.

We know of no systematic empirical study of $G(n)$ in the literature. In the course of verifying that the ERH implies $G(n) \leq 3 \log^2 n$ [1], the first author computed $G(n)$ for $n \leq 10^6$. Brown and Zassenhaus [4] computed $G(p)$ for every prime p less than 10^6, and conjectured that with probability "almost (but not equal to) one", the first $\lfloor \log p \rfloor$ primes will generate $\mathbf{Z}_p^*$. It is not clear how this should be interpreted, but our work does suggest that the natural logarithm in this conjecture should be replaced by a logarithm to the base 2.

Conjectures similar to ours have been made for the least quadratic nonresidue $\bmod p$, when p is prime. For example, Elliott [10] conjectured that the least quadratic nonresidue $\bmod p$ is $O(\log p)^{1+\varepsilon}$, and Montgomery (personal communication) concluded that according to probability theory, the least quadratic nonresidue modulo a prime p ought to be $O(\log p \log\log p)$. Wagon [35] compared the least nonresidue to $2(\log p \log\log p)$.

The idea that $G(n)$ should be $O(\log n \log\log n)$ may therefore be part of the mathematical folklore. Consequently, the main contribution of this paper should be seen as working out the detailed consequences of this idea (in particular, the "correct" constant) and subjecting them to experimental tests.

Artin's conjecture gives an interesting example of probabilistic modeling in number theory; since this bears on our own work, we discuss it here. Artin stated that any number a, not equal to ± 1 or a square, is a primitive root for infinitely many primes. This was based on a naive density argument, which suggested that about 37% of all primes should have a as a primitive root. By examining numerical data, Lehmer and Lehmer [23] showed that the density was not independent of a, and pointed out a possible cause of the error: the events that a is a qth power residue $\bmod p$ are not necessarily independent for all $q|p-1$. This motivated Heilbronn to compute the precise density (see

[37]), and Hooley [19] to show that Artin's density conjecture (in corrected form) follows from the ERH. (For the strongest unconditional result in the direction of Artin's conjecture, see Heath-Brown [18].)

One interesting consequence of the corrected version of Artin's conjecture is that the initial primes are not uniformly distributed in $\mathbf{Z}_p^*$. This leads to the question of whether a naive model of the type we have proposed is sufficient to explain the behavior of $G(n)$. We will return to this point in the final section.

The remainder of this paper is organized as follows. In §2, we discuss the results from probability theory that we will need. This probabilistic theory leads to several conjectures about $\mathbf{Z}_n^*$, which are presented in §3. We discuss lower bounds for $G(n)$ in §4, and the question of its average behavior in §5. Sections 6 and 7 present our experimental methods and numerical data. Finally, the arguments for our conjectures are reviewed in §8.

2. Probabilistic and analytic background

In this section we collect some results for further use, mostly for lack of a suitable reference. We will assume that the reader is familiar with analytic number theory and probability theory. Good references for this material include Davenport [8] and Feller [14].

If A is a finite abelian group, let

$$r(A) = \max\{n : A \text{ is a direct sum of } n \text{ nontrivial cyclic groups}\};$$

we will call this the *group rank* of A. Clearly, then, $r(A) \leq \log_2 |A|$, and the factors occurring in a maximal direct sum are unique up to isomorphism.

A can always be expressed as a direct sum of its p-Sylow subgroups A_p. We let $r_p(A) = r(A_p)$ denote the group rank of its p-Sylow subgroup, and call this the *p-rank*. Each A_p is isomorphic to a direct sum of copies of $\mathbf{Z}_{p^e}$, for various e, and this allows us to think of a set of m elements of A as an $m \times r$ array. Such a set generates A if and only if the subarray of entries chosen from the $\mathbf{Z}_{p^e}$'s has rank r_p, when each entry is reduced $\bmod p$.

We now imagine a process that chooses elements of A at random; we say that the ith stage is complete when the elements chosen so far generate a subgroup of group rank i. The waiting time until A is generated is the sum of the waiting times for each stage. (Note that stages can be skipped.)

Theorem 2.1. *Let T be the number of random samples needed to generate a finite abelian group A. We have*

$$\Pr[T \geq r(A) + x] \leq 2(2 - e^{\lambda})^{-1} e^{-\lambda x},$$

whenever $0 < \lambda < \log 2$.

Proof. The worst case is easily seen to be $A = (\mathbf{Z}_2)^r$. We have $T = \sum_{i=1}^{r} T_i$, where T_i is the time required for the ith stage; note that the T_i are independent geometric random variables. By applying the argument of Chernoff [6] to this sum, we obtain

$$\Pr\left[\sum_{i=1}^{r} T_i \geq r + x\right] \leq e^{-\lambda x} \prod_{i=1}^{r} \frac{p_i}{1 - q_i e^{\lambda}},$$

where $p_i = 1 - 1/2^{r-i+1}$ and $q_i = 1 - p_i$, and $0 < \lambda < \log 2$. The estimate in the theorem follows easily from this. □

To apply this result, we need to estimate the group rank of $\mathbf{Z}_n^*$.

Theorem 2.2. *The group rank of* $\mathbf{Z}_n^*$ *satisfies* $r(\mathbf{Z}_n^*) = O(\frac{\log n}{\log\log\log n})$.

Proof. Let $\omega(n)$ denote the number of distinct prime divisors of n; we also define $\omega(1)$ to be 1. By the Chinese Remainder Theorem and the structure of $\mathbf{Z}_{p^e}^*$, we see that

$$r(\mathbf{Z}_n^*) \le \omega(n) + \sum_{p|n} \omega(p-1) \le \omega(n) + \sum_{p \le \log n} \omega(p-1) + \sum_{\substack{p|n \\ p > \log n}} \omega(p-1).$$

Applying the prime number theorem to each term yields the result. □

We note without proof that the expected time to generate an abelian group A is at most $r(A) + O(1)$. This can easily be sharpened by taking the structure of A into account. Since A is a direct sum of its p-Sylow subgroups, we can imagine that samples from the various A_p are chosen independently in parallel.

Theorem 2.3. *Let* A *be an abelian group, such that* $|A|$ *has* k *prime factors. If* T *is the waiting time to generate* A, *we have*

$$\max\{r_p\} \le E(T) \le \max\{r_p\} + O(\log k).$$

Proof. Let Z_p be the time to generate A_p, less the p-rank $r_p(A)$. Then Z_p is stochastically less than an exponential random variable, that is,

$$\Pr[Z_p \ge x] \le e^{-\alpha x},$$

for some $\alpha > 0$. (See [27] for background on stochastic ordering.) For $x \ge 0$, we therefore have

$$\Pr[\max\{Z_p\} < \alpha^{-1}\log k + x] \ge \left(1 - \frac{e^{-\alpha x}}{k}\right)^k \ge 1 - e^{-\alpha x}.$$

Thus, $\max\{Z_p\}$ is stochastically less than $\alpha^{-1}\log k$ plus an exponential random variable; from this the result follows. □

In studying the average value of $G(n)$, we will need the following analytic result.

Lemma 2.4. *We have*

$$\frac{1}{N}\sum_{n \le N} \omega(\varphi(n)) \sim \frac{(\log\log N)^2}{2}.$$

Proof. We have

$$\sum_{n \le N} \omega(\varphi(n)) = \sum_{n \le N} \sum_{p^e \| n} \omega(\varphi(p^e)) = \sum_{p \le N} \sum_{e \ge 1} \omega((p-1)p^{e-1}) \cdot \#\{n \le N : p^e \| n\}$$
$$\le \sum_{p \le N} \frac{N\omega(p-1)}{p} + \sum_{p \le N} \frac{N(\omega(p-1)+1)}{p(p-1)}.$$

Since $\omega(p-1) = O(\log p)$, the second sum is $O(N)$. By a theorem of Halberstam [17], $\sum_{p \le x} \omega(p-1) \sim x\log\log x/\log x$; we combine this with Stieltjes integration by parts to show the first sum is $N(\log\log N)^2/2 + O(N\log\log N/\log N)$. These bounds easily imply the result. □

3. Primes in residue classes of $\mathbf{Z}_n^*$

Motivated by Dirichlet's theorem, we now adopt the following heuristic model: we assume that the primes $2, 3, 5, \ldots$ lie in residue classes $\bmod n$ that are chosen at random.

To familiarize the reader with this model, we discuss the least prime in an arithmetic progression. Heuristic estimates for this have been advanced before [36, 25]; we will derive these estimates anew using our model. We define

$$P(n) = \min\{x : \text{every residue class of } \mathbf{Z}_n^* \text{ contains a prime } \leq x\}.$$

We can formulate a good guess for the growth rate of $P(n)$, using the following "occupancy" model. Assume that we throw balls into m bins at random. The time until each bin has at least one ball has expected value $m\sum_{i=1}^{m} 1/i$ and has a distribution tightly concentrated around $m\log m$ [13]. We now let the bins be the different residue classes of $\mathbf{Z}_n^*$, so that there are $m = \varphi(n)$ bins. If the primes fell into these classes at random, the expected number of primes needed would be asymptotic to $m\log m \sim \varphi(n)\log n$. (This ignores the $O(\log n/\log\log n)$ primes that divide n, but their effect is negligible.) Using the prime number theorem in the form $p_k \sim k\log k$, we see that $P(n)$ should be close to $\varphi(n)(\log n)^2$. This is Wagstaff's conjecture [36].

In the occupancy problem, the probability that some bin remains empty after r balls are thrown is at most $m(1-1/m)^r \leq me^{-r/m}$. With the choice $m = \varphi(n)$ and $r = (2+\varepsilon)m\log m$, the Borel-Cantelli lemma and the prime number theorem imply we should have

$$\limsup_{n\to\infty} \frac{P(n)}{\varphi(n)(\log n)^2} \leq 2.$$

This was conjectured by McCurley [25], and agrees with numerical data. Wagstaff [36] computed $P(n)$ for $11 \leq n \leq 5\times 10^4$, and found four values of n for which $P(n)/(\varphi(n)\log n\log\varphi(n))$ exceeded 2; the largest value of this ratio was 2.209.

We now apply similar reasoning to make a conjecture about $G(n)$, the largest prime needed to generate $\mathbf{Z}_n^*$. Again, we assume that the initial primes act like random elements of $\mathbf{Z}_n^*$ as far as the group structure is concerned. Under this assumption, Theorem 2.1 implies that for any $\beta > 1/2$, the probability that we need more than $r(\mathbf{Z}_n^*)+x$ primes to generate $\mathbf{Z}_n^*$ is $O(\beta^x)$. Applying Theorem 2.2 and the Borel-Cantelli lemma, we conclude that for any $\varepsilon > 0$, we should need more than $(\frac{1}{\log 2}+\varepsilon)\log n$ primes only finitely often. By the prime number theorem, this is equivalent to the following statement.

Conjecture 1. *We have*

$$\limsup_{n\to\infty} \frac{G(n)}{\log n\log\log n} \leq \frac{1}{\log 2}.$$

We remark that our probabilistic model is oversimplified, but not so much as to affect the above conjecture. For example, we should not consider the primes dividing n as possible elements of $\mathbf{Z}_n^*$, but since n has $O(\log n/\log\log n)$ prime factors, it does no harm to do so. Similarly, for reasons connected with

Artin's conjecture (see §8), we may wish to omit the prime divisors of $\varphi(n)$ from consideration, but these are also few in number.

4. A heuristic lower bound for $G(n)$

In this section we give a different probabilistic argument, which suggests it is unlikely that the constant in Conjecture 1 could be reduced below $1/\log 2$.

Note that if $(p_i|p) = +1$ for $i = 1, \dots, k$, then p splits completely in the number field $K = \mathbf{Q}(\sqrt{2}, \dots, \sqrt{p_k})$. By the Chebotarev density theorem, about $1/2^k$ of the primes will have this property. We now make the heuristic assumption that each prime splits independently with probability $1/2^k$. The probability that more than m primes must be sampled to find one that splits is $(1 - q_k)^m$, where $q_k = 1/2^k$. (Here we are ignoring the first k primes.) By the Borel-Cantelli lemma, then, if p is the least prime splitting in K, we should have $p \leq p_{k2^k}$, with finitely many exceptions. Applying the prime number theorem twice, we obtain the following conjecture.

Conjecture 4.1. *There is an infinite sequence of primes p for which*

$$G(p) \geq \frac{1}{\log 2} \log p \log\log p(1 + o(1)).$$

If we assume the ERH, we get an asymptotic lower bound with a slightly worse constant. This provides a numerical version of a theorem of Montgomery [28].

Theorem 4.2 (ERH). *There is an infinite sequence of primes p for which*

$$G(p) \geq \frac{1}{2\log 2} \log p \log\log p(1 + o(1)).$$

Proof. Let Δ_K denote the absolute value of K's discriminant. According to the effective Chebotarev theorem of Lagarias and Odlyzko [22], the ERH implies the existence of a prime p that splits in K, with $p = O(\log^2 \Delta_K)$. Using bounds for the discriminants of Kummer fields (i.e., Theorem 9.5.2 of [7]) and the prime number theorem, we have $\log \Delta_K = O(k2^k \log k)$. Thus, $\log p \leq 2k \log 2 + O(\log k)$. Since $G(p) > p_k$, this gives the result. □

5. The average value of $G(n)$

In this section we give arguments to support the idea that $\overline{G}(N)$, the average value of $G(n)$ for $n \leq N$, grows like $\log\log N \log\log\log N$. In a paper devoted to heuristics, it is well to provide at least one rigorous theorem.

Theorem 5.1. *Let $\overline{G}(N) = \frac{1}{N}\sum_{n=1}^{N} G(n)$. We have*

$$\overline{G}(N) \geq (1 + o(1)) \log\log N \log\log\log N.$$

Proof. Let $\omega(n)$ denote the number of distinct prime divisors of n. Observing that at least $r_2(\mathbf{Z}_n^*)$ primes not dividing n are needed to generate $\mathbf{Z}_n^*$, we see that $G(n) \geq p_{\omega(n)}$ for all $n \geq 3$. We now temporarily change the definition of $G(n)$ so that this holds for all $n \geq 1$; this will not affect the asymptotic result.

Rosser [30] showed that the kth prime is greater than $k \log k$. The Erdős-Kac theorem [12] implies that $\overline{\omega}(N)$, the mean value of $\omega(n)$ for $n \leq N$, is asymptotic to $\log\log N$. Using these two facts and Jensen's inequality, we

obtain

$$\overline{G}(N) \geq \frac{1}{N}\sum_{n=1}^{N} p_{\omega(n)} \geq \frac{1}{N}\sum_{n=1}^{N} \omega(n)\log\omega(n)$$
$$\geq \overline{\omega}(N)\log\overline{\omega}(N) \sim \log\log N \log\log\log N . \quad \square$$

We now give a heuristic argument for an upper bound that matches Theorem 5.1. Let $G(n) = p_{k(n)}$. We temporarily ignore the primes dividing n, and heuristically consider $k(n)$ for $n \leq N$ as a random variable. If this were true, we could apply Theorem 2.3, observing that the 2-Sylow group has the largest group rank, to obtain

$$\overline{k}(N) = \overline{r}_2(N) + \frac{1}{N}\sum_{n\leq N} O(\log\omega(\varphi(n))) .$$

(Here $\overline{k}(N)$ denotes the mean value of $k(n)$ for $n \leq N$; similarly for $\overline{r}_2(N)$.) Since $r_2(n) = \omega(n) + O(1)$, the mean value of $r_2(n)$ is asymptotic to $\log\log N$. By Jensen's inequality and Lemma 2.4, the mean value of $\log\omega(\varphi(n))$ is $O(\log\log\log N)$. By the Erdős-Kac theorem, therefore, $k(n)$ should be tightly distributed around an asymptotic mean value of $\log\log N$. This suggests the error in ignoring primes dividing n should be negligible, since $\sum_{p<O(\log\log N)} 1/p = O(\log\log\log N)$. We should therefore have

$$\overline{G}(N) = \frac{1}{N}\sum_{n=1}^{N} p_{k(n)} \sim p_{\overline{k}(N)} \leq (1 + o(1))\log\log N \log\log\log N .$$

Combining the two bounds, we get the following conjecture.

Conjecture 5.2. *We have*

$$\overline{G}(N) \sim \log\log N \log\log\log N .$$

6. Computing $G(n)$

This section describes our procedures for computing $G(n)$. In particular, we describe a parallel algorithm that allowed us to find $G(n)$ for each $n \leq 10^9$.

We define a *character* of order p to be a homomorphism from $\mathbf{Z}_n^*$ to the additive group $\{0, \ldots, p-1\}$. (Note that this differs slightly from the usual definition.) We can obtain generators for the character group of $\mathbf{Z}_n^*$ from the factorization of n, using the Chinese Remainder Theorem. This is done in a straightforward way which we will not go into here, except to note that the characters of order 2 can be efficiently computed as Jacobi symbols. Otherwise, evaluation of a character of order p entails the solution of a discrete logarithm problem in a group of order p. We can, however, easily decide if such a character evaluates to 0 or not.

To compute the smallest generating set for $\mathbf{Z}_n^*$, we handle each p-Sylow subgroup separately. This requires us to factor n and further factor $p-1$ for each prime p dividing n. Assuming this is done, we do the following for each p: Choose characters $\chi_1, \ldots, \chi_r$ that generate the character group of the

p-Sylow subgroup. Then find the least x such that the matrix

$$\begin{pmatrix} \chi_1(2) & \chi_2(2) & \cdots & \chi_r(2) \\ \vdots & \vdots & \ddots & \vdots \\ \chi_1(x) & \chi_2(x) & \cdots & \chi_r(x) \end{pmatrix}$$

has rank r. (Here it is convenient to say that a character takes the value 0 when its argument is not a unit $\bmod n$.) $G(n)$ is then the maximum of the x's computed in this fashion.

The above sketch can be filled in to give a sequential algorithm that will compute $G(n)$ for $n \leq N$ in polynomial amortized time. (That is, assuming ERH, the total time is $N(\log N)^{O(1)}$.) The main idea is to use a sieve to factor each $n \leq N$; this guarantees that all relevant factorizations will be available when needed. However, it must also be checked that the discrete logarithm computations will not be burdensome. To do this, we first observe that if p appears only once in the factorization of $\varphi(n)$, it suffices to find the first x with a nonzero character value. Thus, characters of order p need only be evaluated if $p^2|\varphi(n)$. For such p, we have $p \leq \sqrt{N}$; under this restriction and the ERH, it can be shown that $O((\log N)^4/p^2)$ of the $n \leq N$ will have $\varphi(n)$ divisible by p^2. (Naively, one might expect this fraction to be about p^{-2}, but this is false [9].) Even if a brute force search for the discrete logarithm is used, the total cost of all discrete log computations is $\sum_{p<\sqrt{N}}(\log N)^{O(1)}/p$, which is small.

We parallelize the above algorithm by giving each processor the task of computing $G(n)$ for each integer in a block of size k. We can still use the sieve of Eratosthenes to factor the n, provided we give each processor a list of the primes less than $\sqrt{N}$. (This is the parallel segmented wheel sieve, as described by Sorenson and Parberry [31].) However, this leaves the problem of factoring $p-1$; these are simply looked up in a table (if small) and factored by brute force (if large).

We assume available m worker processors $P = \{P_1, \ldots, P_m\}$, and a single master processor M. The latter is responsible for assigning work to an idle P_i, and each P_i is responsible for doing the work assigned to it and for communicating (to M) when this work is complete. It is only necessary for M to send messages to P_i, and for P_i to reply in kind; communication amongst the workers is not necessary.

The master M partitions the interval $\{1, \ldots, N\}$ into subintervals of size k. It is convenient if $k|N$, though in practice the subintervals need not be of identical size. M communicates the bounds of a subinterval to an idle P_i. Although subintervals have equal size, the computation time required for a subinterval varies. Hence, the processors must operate asynchronously. The tradeoff in choosing k is the following: N/k should be large compared to m, so as to equalize the total time spent by each worker. However, if k is too small, communication costs dominate the computation.

To compute record values of $G(n)$ in parallel, M maintains a current record table, R. When assigning work to an idle P_i, then M also sends a copy of R to P_i. A new copy of this is returned, containing any new records that P_i encounters during its computation. M must merge the table received from P_i with R; that is, M replaces superseded records in R with the new records

received from P_i, and inserts new records as necessary.

Any long computation must be fault-tolerant. To accomplish this, M checkpoints the computation every C subintervals. After assigning C subintervals, M waits until the computations of outstanding subintervals complete. Then, M logs the current record table as well as the bounds of the successfully completed interval. If a processor or channel faults, the computation can be restarted from the last checkpoint.

The computation of $G(n)$ for $n \leq 10^9$ was performed on a 64-processor Thinking Machines CM-5. Subintervals were of length $k = 2^{15}$ and the computation was checkpointed every 2^9 subintervals. Each worker was provided with a list of the first 2^{12} primes, together with the factorizations of $p - 1$ in this range. The computation required approximately 54 hours. This time is approximate, since redundant computations were performed (due to faults), and the machine was timeshared with other users.

7. Comparisons with numerical data

In this section, we discuss some experimental results lending support to the heuristic theory.

We computed $G(n)$ for each $n \leq 10^9$. The first two columns of Table 1 give the *record* values of $G(n)$; that is, values of n for which $G(n)$ exceeds

Table 1. Record values of $G(n)$ for $n \leq 10^9$

n	$G(n)$	$a(n)$	$a'(n)$
3	2	13.417211	1.000000
4	3	4.592292	1.000000
6	5	3.316650	1.666667
12	7	2.145161	1.400000
20	11	2.319711	1.571429
24	13	2.452159	1.857143
120	17	1.571711	1.307692
780	23	1.262652	1.000000
920	29	1.533720	1.260870
1364	37	1.797547	1.275862
6090	41	1.506324	1.108108
26220	47	1.380249	1.093023
53570	53	1.412988	1.127660
67044	59	1.528140	1.113208
205608	61	1.380156	1.033898
249690	67	1.482887	1.135593
2225685	73	1.290801	1.000000
3442296	79	1.341717	1.082192
5053620	97	1.591658	1.227848
60369855	101	1.354115	1.041237
191895456	103	1.269674	1.000000
475528443	107	1.239532	1.000000
715236599	109	1.229120	1.000000

all previous values. These values are compared with two predictions from our heuristic theory in the third and fourth columns. We define

$$a(n) = \frac{G(n)\log 2}{\log n \log\log n};$$

the theory predicts that $\limsup a(n) = 1$. This uses the asymptotic value $p_k \sim k\log k$, and is not all that accurate for the small values of $G(n)$ we observed. For this reason, it is better to compare $G(n)$ with p_k, where $k = \lfloor \log_2 n \rfloor$. To do this, we define

$$a'(n) = \frac{G(n)}{p_{\lfloor \log_2 n \rfloor}}.$$

It is interesting to compare the data of Table 1 with the theory of extreme values due to Gumbel [16]. According to this theory, if one observes i.i.d. samples $X_1, X_2, \ldots$ from some distribution, the time between successive record values of X_i is controlled by the tail of their common distribution. In particular, because Theorem 2.1 says that the probability that more than t primes are needed to generate $\mathbf{Z}_n^*$ goes down roughly as 2^{-t}, we might guess that the interval between record values of $G(n)$ should approximately double with each successive record. This is happening toward the end of Table 1. Also, the number of records in this table, 23, is not too far from the "predicted" number of $\log 10^9 + \gamma = 21.300\ldots$.

We also computed the average value of $G(n)$ for $n \le 10^9$ and found it to be 13.032. This is in agreement with our heuristic theory, as we now explain. As a general rule, we have generators for $\mathbf{Z}_n^*$ as soon as we can generate the 2-Sylow subgroup. (Theorem 2.3 and Lemma 2.4 provide an explanation for this, if we note that the 2-Sylow subgroup maximizes the group rank.) The average group rank of the 2-Sylow subgroup for $n \le N$ is $3/8 + \sum_{2<p\le N} 1/p \sim \log\log N + 0.139\ldots$. (Because $\mathbf{Z}_n^*$ is not necessarily cyclic when n is a power of 2, the prime 2 must be treated specially.) The average number of random elements needed to generate $(\mathbf{Z}_2)^r$ is about $r + 1.606\ldots$. Now, a further correction term should be added to account for small prime divisors; for example, 3 divides n a third of the time, and for these n we will need one additional prime. Since $\log\log 10^9 + 0.139\ldots + 1.606\ldots = 4.776\ldots$, it is reasonable to include a correction for the first five primes of $1/2 + 1/3 + 1/5 + 1/7 + 1/11 = 1.267\ldots$. Adding everything up, we predict the average number of primes to be $6.044\ldots$. Because $p_6 = 13$, this agrees well with the computed average of 13.032.

As a further check on the theory, we computed $G(n)$ for the *pseudosquares* tabulated by Lehmer, Lehmer, and Shanks [24] and Stephens and Williams [32]. These are numbers $n \equiv 1 \bmod 8$ with $(n|p) = +1$ for all small p; they are therefore good candidates for unusually large values of $G(n)$. We found in every case that $G(n)$ is the first p with $(n|p) = -1$.

Table 2 compares $G(n)$ against two asymptotic estimates provided by the heuristic theory. We define $a(n)$ as before, and also introduce

$$a''(n) = \frac{G(n)}{p_{\lfloor \log_2(n/8) \rfloor}}.$$

The rationale for $a''(n)$ is that the numbers n are constrained to a residue

TABLE 2. $G(n)$ for pseudosquares

n	$G(n)$	$a(n)$	$a''(n)$
17	3	0.704766	1.500000
73	5	0.554642	1.000000
241	7	0.519767	1.000000
1009	11	0.570000	0.846154
2641	13	0.554055	0.684211
8089	17	0.596046	0.739130
18001	19	0.588953	0.612903
53881	23	0.612721	0.621622
87481	29	0.726419	0.707317
117049	31	0.749358	0.756098
515761	37	0.756709	0.787234
1083289	41	0.777178	0.694915
3206641	43	0.735039	0.704918
3818929	47	0.790763	0.770492
9257329	53	0.825252	0.746479
22000801	59	0.855436	0.808219
48473881	67	0.913315	0.848101
175244281	71	0.880823	0.797753
427733329	79	0.921682	0.814433
898716289	83	0.922162	0.821782
2805544681	101	1.044871	0.943925
10310263441	103	0.986793	0.911504
23616331409	107	0.978531	0.842520
85157610409	113	0.964836	0.824818
196265095009	131	1.071766	0.942446
2871842842801	149	1.072673	0.914110
26250887023729	157	1.026597	0.877095
112434732901969	173	1.066061	0.905759
178936222537081	181	1.095075	0.937824
696161110209049	193	1.108382	0.969849
2854909648103881	197	1.074184	0.833408
6450045516630769	211	1.117676	0.929515
11641399247947921	227	1.177963	0.991266

class mod 8, and it is therefore reasonable to take the number of "samples" to be $n/8$.

It will be observed that Table 1 corresponds better to the heuristic theory than Table 2 does. We have not attempted to explain this, other than to observe that the numbers n in Table 2 were not chosen to be record values of $G(n)$; for this reason, $a''(n)$ might be a bit smaller than 1.

8. CRITIQUE

We may summarize the argument that the true growth rate of $G(n)$ is $O(\log n \log\log n)$ as follows. First, the heuristic model that small primes lie

in random residue classes of $\mathbf{Z}_n^*$ gives numerically accurate predictions, both for the record values and for the average value of $G(n)$. In addition, this model suggests conjectures about primes in residue classes that were derived independently by other means. Second, a "random splitting" argument based on Chebotarev's theorem, and the ERH, suggest that no bound for $G(n)$ below $O(\log n \log\log n)$ can be valid.

Aside from the obvious objection that no randomization is involved for the real primes, there are several others that could be raised against our arguments. The first one is that our theory has assumed sampling with replacement, whereas the primes $2, 3, 5, \ldots$ are obviously distinct. Sampling with replacement, however, will only increase the time needed to construct a generating set, so our upper bound argument is not affected. Another point is that our heuristic theory does not take the effect of primes dividing n into account; one might therefore expect a more complicated stochastic process than the one we have used to be necessary. Finally, it is known that the initial primes do *not* represent random samples from $\mathbf{Z}_n^*$.

In the context of Artin's conjecture, this nonuniformity appears in examples of the following type. Consider a prime p for which 5 is a potential primitive root. If $p \equiv 1 \bmod 5$, we have $(5|p) = +1$ by quadratic reciprocity, so if 5 is not a fifth power $\bmod p$, it cannot be a quadratic nonresidue. Thus, there is a "coupling" between the values taken by 5 in the 2-Sylow group and the 5-Sylow group of $\mathbf{Z}_p^*$; they cannot be thought of as independent.

Using results of Elliott [11], we can express this deviation from randomness in the following way. For an integer $k > 0$, let E_k denote the mean value of the least kth power nonresidue $\bmod p$, where the average is taken over all primes $p \equiv 1 \bmod k$. A naive probabilistic argument suggests that this mean value should equal

$$\widehat{E}_k = \frac{k-1}{\varphi(k)} \sum_{n=1}^{\infty} \frac{p_n}{k^n}.$$

However, the true mean value is given by

$$E_k = \sum_{n=1}^{\infty} p_n \left(\frac{1}{d_{n-1}} - \frac{1}{d_n} \right),$$

where d_n is the degree of the number field $\mathbf{Q}(\zeta_k, \sqrt[k]{p_1}, \ldots, \sqrt[k]{p_n})$. Although these agree when k is prime, they are not the same in general; for example, if $k = 8$, the true value is about 7% higher than the naive probability argument would predict. Because $\widehat{E}_k \leq E_k$, one can think of Elliott's results as reflecting a small, but systematic, bias for small primes to lie in nontrivial subgroups of $\mathbf{Z}_p^*$.

We argue, though, that the nonrandomness reflected in Artin's corrected conjecture and Elliott's results is not great enough to affect our conclusions. In the first place, the effect only applies to primes dividing $p - 1$ (see Lemmas 4 and 5 of [11]), and we can easily compensate for this by banishing the primes dividing $\varphi(n)$ from our model. In the second place, the waiting time to generate $\mathbf{Z}_n^*$ is essentially the waiting time to generate the 2-Sylow subgroup, and we know of no substantial deviations from randomness for quadratic characters.

For example, a result of Baillie and Wagstaff [2, Theorem 9] shows that when n is odd, $L(n)$, the least prime p with $(p|n) \neq 1$, has a mean value in accordance with probability theory, provided one accounts for the primes dividing n. (Dirichlet's class number formula [3, p. 346] can be interpreted as reflecting a bias toward small quadratic residues, but this effect is negligible when n is large.)

For these reasons, we conclude that the simplifications in our model are apparently not drastic enough to affect the growth rate of $G(n)$. Although a more refined model may be worth considering for other reasons, we will not pursue this further here.

In closing, something should be said about why heuristic arguments are worth our attention. After all, this paper presents only one new theorem about $G(n)$. However, many practical algorithms have been designed under the premise that a number-theoretic function behaves randomly. For this reason alone, it is important to state such intuitive ideas in a precise and falsifiable manner. This work will have been justified if it leads to future theorems characterizing the growth rate of the important function $G(n)$, or to new observations indicating that its behavior deviates from the predictions of our simple probabilistic model.

Acknowledgments

The support of the National Science Foundation, via grants CCR-8552596 and CDA-9024618, is gratefully acknowledged. We would also like to thank Hugh Montgomery for useful discussions, and Todd Proebsting for help with sieves.

Bibliography

1. E. Bach, *Explicit bounds for primality testing and related problems*, Math. Comp. **55** (1990), 355–380.
2. R. Baillie and S. S. Wagstaff, Jr., *Lucas pseudoprimes*, Math. Comp. **35** (1980), 1391–1417.
3. Z. I. Borevich and I. R. Shafarevich, *Number theory*, Academic Press, New York, 1966.
4. H. Brown and H. Zassenhaus, *Some empirical observations on primitive roots*, J. Number Theory **3** (1971), 306–309.
5. D. A. Burgess, *On character sums and primitive roots*, Proc. London Math. Soc. (3) **12** (1962), 179–192.
6. H. Chernoff, *A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations*, Ann. Math. Statist. **23** (1952), 493–507.
7. H. Cohn, *A classical invitation to algebraic numbers and class fields*, Springer-Verlag, New York, 1978.
8. H. Davenport, *Multiplicative number theory*, Springer-Verlag, New York, 1980.
9. R. E. Dressler, *A property of the φ and σ_j functions*, Compositio Math. **31** (1975), 115–118.
10. P. D. T. A. Elliott, *The distribution of primitive roots*, Canad. J. Math. **21** (1969), 822–841.
11. ——, *A problem of Erdős concerning power residue sums*, Acta Arith. **13** (1967), 131–149; Corrigendum, ibid. **14** (1968), 437.
12. P. Erdős and M. Kac, *On the Gaussian law of errors in the theory of additive number theoretic functions*, Amer. J. Math. **62** (1940), 738–742.
13. P. Erdős and A. Rényi, *On a classical problem of probability theory*, MTA Mat. Kut. Int. Közl. **6A** (1961), 215–220. Reprinted in *Selected papers of Alfréd Rényi*, vol. 2, Akademiai Kiado, Budapest, 1976, pp. 617–621.

14. W. Feller, *Introduction to probability theory and its applications*, Wiley, New York, 1968.
15. S. Graham and C. J. Ringrose, *Lower bounds for least quadratic nonresidues*, Analytic Number Theory: Proceedings of a Conference in Honor of Paul T. Bateman (B. C. Berndt et al., eds.), Birkhäuser, Boston, 1990, pp. 269–309.
16. E. J. Gumbel, *Statistics of extremes*, Columbia Univ. Press, New York, 1958.
17. H. Halberstam, *On the distribution of additive number theoretic functions*. III, J. London Math. Soc. **31** (1956), 14–27.
18. R. Heath-Brown, *Artin's conjecture for primitive roots*, Quart. J. Math. Oxford Ser. (2) **37** (1986), 27–38.
19. C. Hooley, *On Artin's conjecture*, J. Reine Angew. Math. **225** (1967), 209–220.
20. J. L. Hafner and K. S. McCurley, *A rigorous subexponential algorithm for computation of class groups*, J. Amer. Math. Soc. **4** (1989), 837–850.
21. G. Kolesnik and E. G. Straus, *On the first occurrence of values of a character*, Trans. Amer. Math. Soc. **246** (1978), 385–394.
22. J. C. Lagarias and A. M. Odlyzko, *Effective versions of the Chebotarev density theorem*, Algebraic Number Fields (A. Fröhlich, ed.), Academic Press, London, 1977, pp. 409–464.
23. D. H. Lehmer and E. Lehmer, *Heuristics, anyone?* Studies in Mathematical Analysis and Related Topics (G. Szegö, ed.), Stanford Univ. Press, Stanford, CA, 1962, pp. 202–210.
24. D. H. Lehmer, E. Lehmer, and D. Shanks, *Integer sequences with prescribed quadratic character*, Math. Comp. **24** (1970), 433–451.
25. K. S. McCurley, *The least r-free number in an arithmetic progression*, Trans. Amer. Math. Soc. **293** (1986), 467–475.
26. G. L. Miller, *Riemann's hypothesis and tests for primality*, J. Comput. System Sci. **13** (1976), 300–317.
27. A. W. Marshall and I. Olkin, *Inequalities: theory of majorization and its applications*, Academic Press, New York, 1979.
28. H. L. Montgomery, *Multiplicative number theory*, Lecture Notes in Math., vol. 227, Springer-Verlag, New York, 1971.
29. M. O. Rabin, *Probabilistic algorithm for testing primality*, J. Number Theory **12** (1980), 128–138.
30. J. B. Rosser, *The nth prime is greater than* $n \log n$, Proc. London Math. Soc. **45** (1939), 21–44.
31. J. Sorenson and I. Parberry, *Two fast parallel prime number sieves*, Technical Report CRPDC-91-8, Center for Research in Parallel and Distributed Computing, University of North Texas, July 1991; Inform. Comput. (to appear).
32. A. J. Stephens and H. C. Williams, *An open architecture number sieve*, Number Theory and Cryptography (J. H. Loxton, ed.), Cambridge Univ. Press, Cambridge, 1990, pp. 38–75.
33. G. Tonelli, *Bemerkung über die Auflösung quadratischer Congruenzen*, Gött. Nachr. (1891), 344–346.
34. D. Suryanarayana, *On* $\Delta(x, n) = \varphi(x, n) - x\varphi(n)/n$, Proc. Amer. Math. Soc. **44** (1974), 17–21.
35. S. Wagon, *The evidence: primality testing*, Math. Intelligencer **8** (1986), 58–61.
36. S. S. Wagstaff, Jr., *Greatest of the least primes in arithmetic progressions having a given modulus*, Math. Comp. **33** (1979), 1073–1080.
37. A. E. Western and J. C. P. Miller, *Tables of indices and primitive roots*, Cambridge Univ. Press, Cambridge, 1968.

COMPUTER SCIENCES DEPARTMENT, UNIVERSITY OF WISCONSIN, MADISON, WISCONSIN 53706
E-mail address, E. Bach: bach@cs.wisc.edu
E-mail address, L. Huelsbergen: lorenz@cs.wisc.edu