

# Probability Analysis and Comparison of Well-Known Integer Factorization Algorithms

Duggirala Meher Krishna<sup>1</sup> and Duggirala Ravi<sup>2</sup>

<sup>1,2</sup>*Gayatri Vidya Parishad College of Engineering (Autonomous)  
Madhurawada, Visakhapatnam– 530 048, Andhra Pradesh, India.*

## Abstract

Two prominent methods for integer factorization are those based on general integer sieve and elliptic curve. The general integer sieve method can be specialized to quadratic integer sieve method. In this paper, a probability analysis for the success of these methods is described, under some reasonable conditions. The estimates presented are specialized for the elliptic curve factorization. These methods are compared through heuristic estimates. It is shown that the elliptic curve method is a probabilistic polynomial time algorithm under the assumption of uniform probability distribution for the arising group orders and clearly more likely to succeed, faster asymptotically.

**Keywords:** Integers; Prime numbers; Unique factorization theorem; General integer sieve; Elliptic curve method.

## 1. INTRODUCTION

In this paper, the success probabilities for two prominent methods, *viz.*, general integer sieve method and elliptic curve method, are presented. The estimates are specialized for the elliptic curve factorization algorithm. The random variables studied are (1) the number generated by exponentiating a chosen fixed base random number to various random integer exponents, for general integer sieve method, and (2) the group orders of the elliptic curve groups, with restriction to  $\text{mod } p$ , for each (as yet unknown)

prime factor  $p$  of the integer modulus to be factored. The common assumptions taken in our estimates are that the probabilistic events arising from the consideration of various different smaller prime numbers being factors of any particular realization (sample) of the random variable are mutually independent. With the assumption of independence of events corresponding to divisibility by different smaller prime numbers, the probabilities of success are shown to be fairly optimistic. The general integer sieve needs the random base point to be a group generator (primitive in this sense), which may be difficult to ensure. The merits of elliptic curve method are highlighted, with a caution concerning the widths of the intervals of the possible group orders. Nevertheless, the estimated probabilities of success do not depend too heavily on this fact, as they are applicable to random samples from any arbitrary interval of considerable width, for asymptotic analysis.

## 2. ESTIMATION OF SUCCESS PROBABILITIES

Let  $\mathbb{Z}$  be the ring of integers, and  $\mathbb{N}$  be the set of positive integers. Let  $N$  be a very large positive integer to be factored, and let  $\mathbb{Z}_N$  be the ring of integers with arithmetic operations taken  $\text{mod } N$ .

Let  $L_{\min}, L_{\max} \in \mathbb{Z}$  be such that  $L_{\min} < L_{\max}$  and  $L_{\max} - L_{\min}$  is very large. The consecutive prime numbers are listed in the ascending order as follows:  $2 = q_1, 3 = q_2, 5 = q_3, \dots$ , so that  $q_i$  is the  $i$ -th prime number, for  $i \in \mathbb{N}$ . Let  $k$  be a small positive integer, but still large enough that the asymptotic estimates hold good, and let  $n$  be the largest positive integer, such that  $q_n < \max\{|L_{\min}|, |L_{\max}|\}$ . Let  $X$  be a random variable taking integer values in the interval  $\mathcal{I} = [L_{\min}, L_{\max}]$ , with uniform probability distribution.

**Proposition 1.** *In the notation just discussed, the probability  $\pi_X(z)$  of the event that a sample of the random variable  $X$  is divisible by a positive integer  $z \geq 2$  is approximately  $\frac{1}{z}$ , and more precisely the following bounds hold good:*

$$\frac{1}{z} - \frac{1}{L_{\max} - L_{\min}} \leq \pi_X(z) \leq \frac{1}{z} + \frac{1}{L_{\max} - L_{\min}} \quad (1)$$

*Proof.* For every positive integer  $z \geq 2$ , the number of integer multiples of  $z$  in  $\mathcal{I}$  are between  $(\frac{L_{\max} - L_{\min}}{z} - 1)$  and  $(\frac{L_{\max} - L_{\min}}{z} + 1)$ . Thus, the probability that a random sample of  $X$  is divisible by  $z$  is between  $\frac{1}{z} - \frac{1}{L_{\max} - L_{\min}}$  and  $\frac{1}{z} + \frac{1}{L_{\max} - L_{\min}}$ , which justifies the assumptions, with appropriate choices of  $z$ .  $\square$

The conjunct consideration concerning the divergence of  $\sum_i \frac{1}{q_i}$  and the convergence of  $\sum_i \frac{1}{q_i^2}$  necessitates taking product spaces. Moreover, the estimates are presented only for elliptic curve factorization algorithm.

## 2.1. Success of Elliptic Curve Factorization

Let  $r = \left\lceil \frac{\log(N)}{\log(q_k)} \right\rceil$ , where the choice of  $k$ , the number of smaller prime factors to be used, is assumed to be considerably larger than 2, such as about 1000. Actually,  $k$  can run into tens of thousands, for practical purposes, and constrained by the condition that  $q_k^r \geq N$ . If  $q_k$  is too small, then  $r$  can be so large that the estimated failure probabilities may become irrelevant. Let  $\mathcal{C}_l(\mathbb{Z}_N)$  be elliptic curves, defined over  $\mathbb{Z}_N$ , for  $1 \leq l \leq r$ . Let  $p$  be a large but unknown prime integer factor  $N$ , such that  $p \leq \sqrt{N}$ , and  $\mathcal{C}_l(\mathbb{Z}_p)$  be the corresponding elliptic curves restricted to  $\mathbb{Z}_p$ , for  $1 \leq l \leq r$ . The group order of  $\mathcal{C}_l(\mathbb{Z}_p)$  is  $p + 1 - a_l$ , where  $-2\sqrt{p} \leq a_l \leq 2\sqrt{p}$ , by Hasse-Weil bounds for the elliptic curve group orders. The probability distribution of  $p + 1 - t$  of the group order of  $\mathcal{C}(\mathbb{Z}_p)$ , as obtained by taking mod  $p$  restriction of a randomly generated elliptic curve  $\mathcal{C}(\mathbb{Z}_N)$  is assumed to be uniform over the interval  $\mathcal{I} = [(\sqrt{p} - 1)^2, (\sqrt{p} + 1)^2]$ .

**Proposition 2.** *Let  $\mathcal{C}_l(\mathbb{Z}_N)$ , for  $1 \leq l \leq r + 2$ , be any  $(r + 2)$  independent samples of the elliptic curves, and  $p$  be a fixed (though unknown yet) prime factor of  $N$ , such that  $p \leq \sqrt{N}$ . Let  $\mathcal{E}_{k+1}$  be the random event that each of the  $(r + 2)$  group orders  $p + 1 - a_l$  of the elliptic curves  $\mathcal{C}_l(\mathbb{Z}_p)$ , for  $1 \leq l \leq r + 2$ , is divisible by a prime factor at least as large as  $q_{k+1}$ , where the prime number  $p$  is assumed to be such that  $p \mid N$  and  $p \geq q_{k+1}$ . Then,  $Pr(\mathcal{E}_{k+1}) \leq \frac{(r+2)(r+1)+8}{2 \times 4 \times (q_{k+1}-1)} + \mathcal{O}\left(\frac{(r+2)(r+1)}{8} \times \frac{\log(\log(p))}{\sqrt{p}}\right)$ . Further, if the approximation  $q_i \approx i \log(i)$ , for sufficiently large positive integer  $i$ , is permitted, then  $Pr(\mathcal{E}_{k+1}) \leq \frac{(r+2)(r+1)+8}{2 \times 4 \times k \times (\log(k+1))^2} + \mathcal{O}\left(\frac{(r+2)(r+1)}{8} \times \frac{\log(\log(p))}{\sqrt{p}}\right)$ .*

*Proof.* Before proceeding with the proof, a justification for the validity of the approximation in the last part is as follows: by the prime number theorem,  $i \approx \frac{q_i}{\log(q_i)} < \frac{q_i}{\log(i)}$ , and  $q_i$  is likely to be larger than  $i \log(i)$ . It may also be noticed that  $\frac{(r+2)(r+1)+8}{8k(\log(k+1))^2} \approx \frac{(r+2)(r+1)+8}{8q_k(\log(k+1))}$ .

The random event  $\mathcal{E}_{k+1}$  in the statement is broken up into the following two parts:  $\mathcal{E}_{k+1} \subseteq E_{k+1,1} \cup E_{k+1,2}$ , where

1.  $E_{k+1,1}$  is the event that there are distinct prime numbers  $q_{i_l} \geq q_{k+1}$ , for  $1 \leq l \leq r + 2$ , such that  $q_{i_l} \mid (p + 1 - a_l)$  and  $q_{i_{l'}} \nmid (p + 1 - a_{l'})$ , for  $l' \neq l$  and  $1 \leq l, l' \leq r + 2$ , and
2.  $E_{k+1,2}$  is the event that there is a prime number  $q_i \geq q_{k+1}$ , such that  $q_i \mid (p + 1 - a_l)$  and  $q_i \mid (p + 1 - a_{l'})$ , for two indexes  $l$  and  $l'$ ,  $l' \neq l$ , where  $1 \leq l, l' \leq r + 2$ .

The two events listed above are not mutually exclusive, but an upper bound for the sum of their probabilities is found, as an estimate for the upper bound of the event in the

statement.

**Part (1).** For the event  $E_{k+1,1}$ , it is observed that, from the simultaneous congruence relations  $p+1 \equiv a_l \pmod{q_{i_l}}$ , for  $1 \leq l \leq r+2$ , the fixed number  $p+1$  can be recovered by the Chinese remainder theorem. The mapping  $a_l \mapsto a_l \pmod{q_{i_l}}$ , for  $1 \leq l \leq r+2$ , induces the homomorphism  $(a_1, \dots, a_{r+2}) \mapsto (a_1 \pmod{q_{i_1}}, \dots, a_{r+2} \pmod{q_{i_{r+2}}})$ , that preserves the algebraic structure. In the proof, it is assumed that the probability distributions remain uniform under the mapping  $a_l \mapsto a_l \pmod{q_{i_l}}$ , for  $1 \leq l \leq r+2$ , with restriction on the domain of possible values of  $(a_1 \pmod{q_{i_1}}, \dots, a_{r+2} \pmod{q_{i_{r+2}}})$ .

By the mutual independence of  $a_l$ , for  $1 \leq i \leq r+2$ , there are at least  $4^{r+2} \prod_{l=1}^{r+2} \sqrt{q_{i_l}}$  many possibilities, in all, for the set of possible realizations  $(a_1 \pmod{q_{i_1}}, \dots, a_{r+2} \pmod{q_{i_{r+2}}})$ , after taking into account the restriction that  $|a_l| \leq 2\sqrt{p}$ . The fixed number  $p+1$  must belong to the set of positive integers that can be reconstructed by any realization of  $(a_1 \pmod{q_{i_1}}, \dots, a_{r+2} \pmod{q_{i_{r+2}}})$ , with  $p$  constrained to be a prime number. Now, the number of possibilities for the realizations for  $(a_1 \pmod{q_{i_1}}, \dots, a_{r+2} \pmod{q_{i_{r+2}}})$ , that could result in the reconstruction of  $p+1$ , with  $p$  restricted to be a prime number at most  $\sqrt{N}$  (or of bit size at most  $\frac{\log_2(N)}{2}$ ), is smaller than  $\prod_{l=1}^r \sqrt{q_{i_l}}$ , because  $(\sqrt{q_k})^r \geq \sqrt{N} > \frac{p+1}{2}$ . Thus,  $Pr(T_{k+1}) \leq \frac{1}{\sqrt{q_{i_{r+1}q_{i_{r+2}}}}} \leq \frac{1}{q_{k+1}}$ . A justification for this approach is given in a separate paragraph following the proof of the second part.

**Part (2).** For the event  $E_{k+1,2}$ , a slightly weaker proof is given in this paragraph, and a more accurate proof is given the correction part below. The event that a prime number  $q_i \geq q_{k+1}$ , such that  $q_i$  divides the group orders of both  $C_l(\mathbb{Z}_N)$  and  $C_{l'}(\mathbb{Z}_N)$ , for some  $l$  and  $l'$ ,  $l \neq l'$  and  $1 \leq l, l' \leq r+2$ , occurs with probability  $\frac{(r+2)(r+1)}{2q_i^2}$ , for any  $i$ , where  $i \geq k+1$ . This probability also accounts for the possibility that  $q_i \mid p+1 - a_l$  and  $q_i \mid p+1 - a_{l'}$ , in case  $a_l = a_{l'}$ , but  $l \neq l'$ , where  $1 \leq l, l' \leq r+2$ , for some prime number  $p \mid N$  and  $p \geq q_{k+1}$ . However, there are at least four possibilities that  $q_i$  divides either component of the pairs  $(p+1 - a_l, p+1 - a_{l'})$ ,  $(p'+1 - a'_l, p+1 - a_{l'})$ ,  $(p+1 - a_l, p'+1 - a'_l)$  and  $(p'+1 - a'_l, p'+1 - a'_{l'})$ , for two distinct prime factors  $p$  and  $p'$  of the composite number  $N$ , of which only one possibility is taken into account, for a fixed  $p$ . Thus, a multiplier by at most the fraction  $\frac{1}{4}$  must be applied. Now,  $\sum_{i \geq k+1} \frac{1}{q_i^2} < \sum_{i \geq k+1} \left[ \frac{1}{q_i - 1} - \frac{1}{q_i} \right] < \frac{1}{q_{k+1} - 1}$ . The result follows by adding it to probability bound in the first part.

If the approximation  $q_i \approx i \log(i)$  is permitted, the probability bound in the second part is as follow:  $\sum_{i \geq k+1} \frac{1}{q_i^2} \approx \sum_{i \geq k+1} \frac{1}{i^2 (\log(i))^2} < \frac{1}{(\log(k+1))^2} \sum_{i \geq k+1} \frac{1}{i^2} < \frac{1}{(\log(k+1))^2} \sum_{i \geq k+1} \left[ \frac{1}{i-1} - \frac{1}{i} \right] < \frac{1}{k(\log(k+1))^2}$ .  $\square$

In the following, a justification for the upper bound for  $Pr(E_{k+1,1})$  and a small correction to the upper bound for  $Pr(E_{k+1,2})$ , assuming that  $N$  is a random integer modulus of a prescribed bit size, are given.

Justification for Upper Bound for  $Pr(E_{k+1,1})$ . Conditional and joint probabilities over the possible random modulus integer  $N$ , of bit size equal to a prescribed parameter  $(\lceil \log_2(N) \rceil)$ , for independent realizations of the tuples  $(a_1, \dots, a_{r+2})$ , with appropriate restrictions on the domains of possible values, are taken into consideration. Let the sequences  $(i_1, \dots, i_{r+2})$ , for  $i_l \neq i_{l'}$  and  $k+1 \leq i_l, i_{l'} \leq n$ , where  $1 \leq l, l' \leq r+2$ ,  $l \neq l'$  and  $n$  is the largest positive integer such that  $q_n \leq (N^{\frac{1}{4}} + 1)^2$ , be enumerated in some particular total order, denoted by  $\prec$ . Let  $X_{(i_1, \dots, i_{r+2})}$  be the event that the group order of  $\mathcal{C}_l(\mathbb{Z}_N)$  is divisible by  $q_{i_l}$ , for  $1 \leq l \leq r+2$ , over all possible integer moduli of bit size  $(\lceil \log_2(N) \rceil)$ , excluding the events  $X_{(j_1, \dots, j_{r+2})}$ , for  $(j_1, \dots, j_{r+2}) \prec (i_1, \dots, i_{r+2})$ , if any. Now

$$\begin{aligned}
 Pr(E_{k+1,1}) &\leq \sum_{(i_1, \dots, i_{r+2})} [ Pr(X_{(i_1, \dots, i_{r+2})}) \times \\
 &\quad Pr(\text{the event that } p \text{ is a large prime number} \\
 &\quad \text{of bit size at most } \frac{\log_2(N)}{2}, \text{ such that,} \\
 &\quad \text{for every } l, q_{i_l} \mid p+1-a_l, \text{ and} \\
 &\quad \text{for some } l', q_{j_{l'}} \nmid p+1-a_{l'}, \text{ whenever} \\
 &\quad (j_1, \dots, j_{r+2}) \prec (i_1, \dots, i_{r+2}), \\
 &\quad \text{where } 1 \leq l, l' \leq r+2 ) ] \\
 &\leq \sum_{(i_1, \dots, i_{r+2})} [ Pr(X_{(i_1, \dots, i_{r+2})}) \times \\
 &\quad Pr(\text{the event that } p \text{ is a large prime number} \\
 &\quad \text{of bit size at most } \frac{\log_2(N)}{2}, \text{ such that,} \\
 &\quad \text{for every } l, q_{i_l} \mid p+1-a_l, \\
 &\quad \text{where } 1 \leq l \leq r+2 ) ] \\
 &\leq \sum_{(i_1, \dots, i_{r+2})} Pr(X_{(i_1, \dots, i_{r+2})}) \times \frac{1}{q_{k+1}} \leq \frac{1}{q_{k+1}}
 \end{aligned}$$

**Small Correction of Upper Bound for  $Pr(E_{k+1,2})$ .** Taking the upper estimate  $\frac{1}{q_i} + \frac{1}{4\sqrt{p}}$  in place of  $\frac{1}{q_i}$ , for  $k+1 \leq i \leq n$ , the following is obtained:

$$Pr(E_{k+1,2}) \leq \sum_{i=k+1}^n \left( \frac{1}{q_i} + \frac{1}{4\sqrt{p}} \right)^2 = \sum_{i=k+1}^n \left( \frac{1}{q_i^2} + \frac{1}{8q_i\sqrt{p}} + \frac{1}{16p} \right)$$

where  $n$  is constrained to be the largest positive integer such that  $q_n$  may possibly divide both  $p + 1 - a$  and  $p + 1 - a'$ , for some  $-2\sqrt{p} \leq a, a' \leq 2\sqrt{p}$ . Since  $\gcd(p + 1 - a, p + 1 - a')$  must divide  $|a - a'| \leq 4\sqrt{p}$ , it may be assumed that  $n \leq \frac{4\sqrt{p}}{\log(4\sqrt{p})}$ , when  $a \neq a'$ . The terms accrued from

1. the sum  $\frac{1}{\sqrt{p}} \sum_{i=k+1}^n \frac{1}{q_i}$ , which can be replaced with  $\frac{\log(\log(q_n))}{\sqrt{p}} \approx \frac{\log(2\log(\sqrt{p}+1))}{\sqrt{p}}$ ;
2. the event that  $a = a'$ , which is  $\frac{1}{4\sqrt{p}}$ , for independent samples  $a$  and  $a'$ , assuming values from the interval  $[-2\sqrt{p}, 2\sqrt{p}]$ ; and
3. the sum  $\sum_{i=k+1}^n \frac{1}{p}$ , which can be replaced with  $\frac{(4\sqrt{p})}{p \log(4\sqrt{p})} = \frac{4}{\sqrt{p} \log(4\sqrt{p})}$

are insignificant for large  $p$ . In the statement of the proposition, the effect of the correction terms is reflected in the addend  $O\left(\frac{(r+2)(r+1)}{8} \times \frac{\log(\log(p))}{\sqrt{p}}\right)$ .

The methods for justification and correction terms are similar to *a priori* and *a posteriori* estimation of the probabilities. This indirect approach is necessitated by the difficulties arising out of the need to deal with the principle of inclusion-and-exclusion in the estimation of the probability of union of events, from the probabilities of independent individual atomic events. For instance, if  $Pr(\mathcal{E}_{k+1})$  is replaced with something like  $\frac{\sum_{i=k+1}^n \frac{1}{q_i}}{\sum_{i=1}^n \frac{1}{q_i}}$ , for some large enough  $n$ , the resulting failure probability may become totally unrealistic. If hyperelliptic curve method can be adapted for factorization, the success probability may hopefully become better.

### 3. COMPARISON WITH GENERAL INTEGER SIEVE FACTORIZATION

Let  $N$  be a large composite positive integer, and  $g \in \mathbb{Z}_N^*$ , where  $\mathbb{Z}_N^*$  is the group of invertible elements  $\bmod N$ , with respect to the multiplication  $\bmod N$ . For a randomly chosen  $t \in \mathbb{Z}_N$ , the event that every prime factor of  $g^t \bmod N$  is at most  $q_k$ . Let  $d_j$  be the discrete logarithm of  $q_j$ , assuming that  $q_j$  belong to the cyclic subgroup generated by  $g$ , for  $1 \leq j \leq k$ . After collecting sufficient number of samples, a system linear equations of the form  $\sum_{j=1}^k \nu_{i,j} d_j \equiv t_i \bmod \phi(N)$  is formed, for  $1 \leq i \leq k$ , where  $\phi(N)$  is the Euler function of  $N$ , which is the group order of  $\mathbb{Z}_N^*$ . Any such relation arise as a result of the factorization  $g^{t_i} = \prod_{j=1}^k q_j^{\nu_{i,j}}$ , for some random samples  $t_i$ , for  $1 \leq i \leq k$ .

From every new relation  $\sum_{j=1}^k \nu_{k+l,j} d_j \equiv t_{k+l} \bmod \phi(N)$ , a vector, consisting of integers  $\tau_{k+l,i}$ ,  $1 \leq i \leq k$ , as components, may be hopefully found, such that  $\sum_{i=1}^k \tau_{k+l,i} \nu_{i,j} \equiv 0 \bmod \phi(N)$ , for  $l = 1, 2, 3, \dots$ . Some of the relations may be redundant, leading to trivial relations. In fact, if two linearly independent relations

$\sum_{j=1}^k \nu_{i,j} d_j \equiv t_i \pmod{\phi(N)}$ , for  $i = 1$  and  $2$ , are obtained, then a linear relation of the form  $\sum_{j=1}^k c_j d_j \equiv 0 \pmod{\phi(N)}$ , for some integers  $c_j$ ,  $1 \leq j \leq k$ , not all 0, can be found. In addition, if  $\rho \mid c_j$ ,  $1 \leq j \leq k$ , for some integer  $\rho \geq 2$ , then a relation of the form  $h^\rho = 1 \pmod{N}$ , for some  $h \in \mathbb{Z}_N^*$ , can be found out. Linear relations, like  $\sum_{j=1}^k c_j d_j \equiv 0 \pmod{\phi(N)}$ , are called trivial, if it so happens that  $\sum_{j=1}^k c_j d_j = 0$ , even without applying  $\pmod{\phi(N)}$ . For quadratic integer sieve,  $\pmod{2}$  restriction (which can be interpreted as the situation corresponding to  $\rho = 2$ ) is taken, with a view to improve the efficiency, because if  $g^{2t} = 1 \pmod{N}$ , for some integer  $t$ , then, with  $h = g^t$ ,  $(h - 1)$  and  $(h + 1)$  may yield nontrivial factors of  $N$  by  $\text{gcd}$ .

The estimation of probability of generating a linear relation in  $d_j$ , for  $1 \leq j \leq k$ , does not carry over from elliptic curve method to general integer sieve, as the term  $(p + 1)$  plays a pivotal role in our estimation of error probabilities of the elliptic curve factorization method. Subsequently, the merits of elliptic curve factorization method are described.

### Merits of Elliptic Curve Factorization

1. the method is probabilistic polynomial time algorithm under the assumption of uniform probability of the group orders for random modulus of given size ;
2. the space requirement is quite small, compared to integer sieve method ;
3. if at least one sample of  $k$ -smooth group order is realized, then the factorization produces a result ; and
4. it is not necessary to assume that the initial random point for any selected curve is a group generator

However, diligence must be exercised while exponentiating by a prime number  $q_i$ , in that the exponentiation may be conducted for at most  $\frac{\log(N)}{2\log(q_i)}$  times, for every positive integer  $i \leq k$ . The number of curve samples also plays an important role, which must be taken in parallel, for each exponentiation by  $q_i$ ,  $1 \leq i \leq k$ .

## 4. CONCLUSION

The probability analysis for the elliptic curve factorization is presented. The method is shown to be a probabilistic polynomial time algorithm, under reasonable assumptions on the probability distribution of the group orders that arise, when restriction to a fixed (but unknown) smaller prime factor of the modulus integer to be factored is taken. The integer modulus to be factored is treated as a random variable of fixed size, because it is

an input to the factorization algorithm. The analysis takes into account the *a priori* and *a posteriori* probabilities. The probability of successful factorization is fairly optimistic.

## **REFERENCES**

- [1] N. Koblitz, “A Course in Number Theory and Cryptography”, *Springer-Verlag*, 1994
- [2] L. Washington, “Elliptic Curves : Number Theory and Cryptography”, *Chapman & Hall / CRC Press*, Taylor & Francis Group, 2008