

# On asymptotic behavior of composite integers $n = pq$

Yasufumi Hashimoto

Received on March 12, 2009

**Abstract.** In this paper, we study the asymptotic behavior of the number of composite integers written by products of two primes. Such integers are sometimes called by the RSA integers, because these are used in the RSA cryptosystems. The number of all such integers has been already studied by Landau, Sathe, Selberg etc. Furthermore, the number of integers with  $n = pq$  and  $p < q < cp$  for a fixed  $c > 1$  was recently studied by Decker and Moree. The aim of this paper is to extend Decker-Moree’s result, and the main theorem describes the asymptotic formula of the number of integers with  $p < q < f(p)$  for a fixed increasing function  $f$ .

*Keywords.* composite integer  $n = pq$ , prime number theorem, RSA cryptosystem

## 1. INTRODUCTION

It is well known that

$$\#\{p: \text{prime}, p < x\} \sim \frac{x}{\log x} \quad \text{as } x \rightarrow \infty.$$

The asymptotic formula above is called by the prime number theorem. For the number of composite integers with  $r (\geq 1)$  distinct prime factors, Landau [10] proved the following asymptotic formula (see also [4]).

$$\begin{aligned} & \#\{n = p_1 \cdots p_r < x \mid p_1, \dots, p_r: \text{primes}\} \\ & \sim \frac{(\log \log x)^{r-1}}{(r-1)!} \frac{x}{\log x} \quad \text{as } x \rightarrow \infty. \end{aligned}$$

Note that this asymptotic formula has been improved by Sathe [12], Selberg [13], Hensley [5] and Hildebrand-Tenenbaum [6].

In this paper, we study the distribution of composite integers  $n = pq$  with two primes  $p, q$ . Such integers are sometimes called by the RSA integers because these are used in the RSA cryptosystem [11] whose security is based on the difficulty to factor  $n$ . In general, it is not easy to factor huge composite integers feasibly without quantum computers (see [15]). However, it is known that the RSA is weak when  $p, q$  satisfy some special conditions. One of such conditions is for the difference between  $p$  and  $q$ . In fact, the computational task of the Fermat factoring algorithm depends on the difference  $|p - q|$  (see, e.g. [8]). Also, Weger [16] found that the secret key in the RSA cryptosystem should be larger as the difference  $|p - q|$  is smaller. Conversely, when one of  $p, q$  is much larger than the other (RSA with such  $p, q$  is called by the unbalanced RSA, see [14]), Boneh-Durfee [1] pointed out that the secret key should be large enough. In this sense, it is important to study

the number of composite integers  $n = pq$  satisfying some conditions of  $p$  and  $q$  for the practical use of the prime numbers. In fact, the following asymptotic formula has been experimentally known among the cryptologists.

$$(1) \quad \#\{n = pq \mid p < q < cp, n < x\} \sim \alpha \frac{x}{(\log x)^2} \quad \text{as } x \rightarrow \infty,$$

where  $\alpha > 0$  is a constant depending on  $c > 1$ . Recently, Decker and Moree [2] proved (1) pure mathematically and found that  $\alpha = 2 \log c$ . In the present paper, we obtain the following result as an extension of the work in [2].

**Theorem 1.** *Let  $f, g : \mathbb{R}_{>1} \rightarrow \mathbb{R}_{>1}$  be increasing functions such that  $f(x) > x$  and  $g(x)f(g(x)) = x$ . Then we have*

$$\begin{aligned} & \#\{n = pq \mid p, q: \text{primes}, p < q < f(p), n < x\} \\ & \sim \begin{cases} \frac{x}{\log x} \log \left( \frac{\log x}{\log g(x)} - 1 \right), & (f(x) \gg x^M \text{ for } \forall M > 0), \\ (\log l) \frac{x}{\log x}, & (f(x) \sim cx^l \text{ for } l > 1, c > 0), \\ (2 \log c) \frac{x}{(\log x)^2}, & (f(x) \sim cx \text{ for } c > 1), \\ 2c \frac{3 - \delta}{1 + \delta} \frac{x^{\frac{\delta+1}{2}}}{(\log x)^2}, & (f(x) - x \sim cx^\delta \\ & \text{for } c > 0, 1/2 < \delta < 1, \\ & \text{and RH holds}), \end{cases} \end{aligned}$$

as  $x \rightarrow \infty$  (where RH is the Riemann hypothesis).

In this paper, we avoid the case where  $f(x) - x = o(x^{1/2})$  because the estimation of the error terms (written by  $A_2$  and  $B$  in the proof of Theorem 1) is difficult. However, the distribution of  $n$  for such  $f(x)$  is important in the analytic number theory. In fact, it relates to the problem to count prime numbers in short intervals. Especially, when

$f(x) = x + (\text{constant})$ , our problem is almost same to the famous prime-pair problem studied by Hardy and Littlewood [3]. Although there are experimental results (see also [9] for the recent research) and conjectures for the prime-pair problem, it still remains as an unsolved problem at the present time.

## 2. PROOF OF THEOREM 1

Let

$$\pi(x) := \#\{p: \text{prime} \mid p < x\} = \sum_{p < x} 1,$$

$$\psi(x) := \sum_{p < x} \log p.$$

It is known that

$$\pi(x) = \text{li}(x) + R_1(x),$$

$$\psi(x) = x + R_2(x),$$

where  $\text{li}(x) := \int_2^x (\log t)^{-1} dt \sim x/\log x$  as  $x \rightarrow \infty$  and the reminder terms  $R_1(x)$  and  $R_2(x)$  are as follows.

$$R_1(x), R_2(x) = \begin{cases} O(xe^{-c(\log x)^{1/2}}), & (\text{unconditionally}), \\ O(x^{1/2+\epsilon}), & (\text{if RH holds}). \end{cases}$$

Note that there are sharper estimates of the reminder terms for the unconditional case (see, e.g. [7]). However, we do not use them in this paper.

Put

$$\pi_{2,f}(x) := \sum_{\substack{p, q: \text{prime} \\ p < q < f(p) \\ pq < x}} 1, \quad \psi_{2,f}(x) := \sum_{\substack{p, q: \text{prime} \\ p < q < f(p) \\ pq < x}} \log(pq).$$

We now estimate  $\psi_{2,f}(x)$  to prove Theorem 1.

We see that

$$\begin{aligned} \psi_{2,f}(x) &= \sum_{p < g(x)} (\log p)\pi(f(p)) + \sum_{p < g(x)} \psi(f(p)) \\ &+ \sum_{g(x) \leq p < x^{1/2}} (\log p)\pi\left(\frac{x}{p}\right) + \sum_{g(x) \leq p < x^{1/2}} \psi\left(\frac{x}{p}\right) \\ &- \sum_{p < x^{1/2}} (\log p)\pi(p) - \sum_{p < x^{1/2}} \psi(p). \end{aligned}$$

Divide the above by  $\psi_{2,f}(x) = A + B$ , where

$$\begin{aligned} A &:= \sum_{p < g(x)} (\log p)\text{li}(f(p)) + \sum_{p < g(x)} f(p) \\ &+ \sum_{g(x) \leq p < x^{1/2}} (\log p)\text{li}\left(\frac{x}{p}\right) + \sum_{g(x) \leq p < x^{1/2}} \frac{x}{p} \\ &- \sum_{p < x^{1/2}} (\log p)\text{li}(p) - \sum_{p < x^{1/2}} p, \end{aligned}$$

$$\begin{aligned} B &:= \sum_{p < g(x)} (\log p)R_1(f(p)) + \sum_{p < g(x)} R_2(f(p)) \\ &+ \sum_{g(x) \leq p < x^{1/2}} (\log p)R_1\left(\frac{x}{p}\right) + \sum_{g(x) \leq p < x^{1/2}} R_2\left(\frac{x}{p}\right) \\ &- \sum_{p < x^{1/2}} (\log p)R_1(p) - \sum_{p < x^{1/2}} R_2(p). \end{aligned}$$

We furthermore divide  $A$  by  $A = A_1 + A_2$ , where

$$\begin{aligned} A_1 &:= \int_2^{g(x)} \text{li}(f(t))dt + \int_2^{g(x)} \frac{f(t)}{\log t} dt \\ &+ \int_{g(x)}^{x^{1/2}} \text{li}\left(\frac{x}{t}\right)dt + \int_{g(x)}^{x^{1/2}} \frac{x}{t \log t} dt \\ &- \int_2^{x^{1/2}} \text{li}(t)dt - \int_2^{x^{1/2}} \frac{t}{\log t} dt, \\ A_2 &:= \int_2^{g(x)} \text{li}(f(t))dR_2(t) + \int_2^{g(x)} f(t)dR_1(t) \\ &+ \int_{g(x)}^{x^{1/2}} \text{li}\left(\frac{x}{t}\right)dR_2(t) + \int_{g(x)}^{x^{1/2}} \frac{x}{t}dR_1(t) \\ &- \int_2^{x^{1/2}} \text{li}(t)dR_2(t) - \int_2^{x^{1/2}} tdR_1(t) \\ &= - \int_2^{g(x)} f'(t)\left(\frac{R_2(t)}{\log f(t)} + R_1(t)\right)dt \\ &+ \int_{g(x)}^{x^{1/2}} \frac{x}{t^2}\left(\frac{R_2(t)}{\log x - \log t} + R_1(t)\right)dt \\ &+ \int_2^{x^{1/2}} \left(\frac{R_2(t)}{\log t} + R_1(t)\right)dt. \end{aligned}$$

We now start estimating  $A_1, A_2$  and  $B$ .

### 2.1. ESTIMATE OF $A_1$ .

It is easy to see that

$$\begin{aligned} \int_{g(x)}^{x^{1/2}} \text{li}\left(\frac{x}{t}\right)dt &= \left[ t\text{li}\left(\frac{x}{t}\right) - x \log \log \left(\frac{x}{t}\right) \right]_{g(x)}^{x^{1/2}} \\ &= x^{1/2}\text{li}(x^{1/2}) - x \log \log(x^{1/2}) \\ &\quad - g(x)\text{li}\left(\frac{x}{g(x)}\right) + x \log \log \left(\frac{x}{g(x)}\right), \\ \int_{g(x)}^{x^{1/2}} \frac{x}{t \log t} dt &= \left[ x \log \log t \right]_{g(x)}^{x^{1/2}} \\ &= x \log \log(x^{1/2}) - x \log \log g(x), \end{aligned}$$

$$\int_2^{x^{1/2}} \left(\text{li}(t) + \frac{t}{\log t}\right)dt = [t\text{li}(t)]_2^{x^{1/2}} = x^{1/2}\text{li}(x^{1/2}).$$

Then we have

$$\begin{aligned} A_1 &= \int_2^{g(x)} \left(\text{li}(f(t)) + \frac{f(t)}{\log t}\right)dt + x \log \left(\frac{\log x}{\log g(x)} - 1\right) \\ &\quad - g(x)\text{li}\left(\frac{x}{g(x)}\right) \\ &=: A_{11} + A_{12} - A_{13}. \end{aligned}$$

We estimate  $A_{11}, A_{12}$  and  $A_{13}$  in the above with the conditions of the growth of  $f$ .

### 2.1.1. The case of $f(x) \gg x^M$ for any $M > 0$ .

For  $A_{11}$  and  $A_{13}$ , we see that

$$\begin{aligned} A_{11} &= \int_2^{g(x)} \left( \text{li}(f(t)) + \frac{f(t)}{\log t} \right) dt \ll 2g(x)f(g(x)) = 2x, \\ A_{13} &= g(x) \text{li}\left(\frac{x}{g(x)}\right) \ll g(x) \left(\frac{x}{g(x)}\right) = x. \end{aligned}$$

Since  $g(x) \ll x^{1/(M+1)}$  for  $f(x) \gg x^M$ , we have

$$A_{12} = x \log \left( \frac{\log x}{\log g(x)} - 1 \right) \gg x \log M,$$

for any  $M > 0$ . Thus we obtain

$$A_1 \sim A_{12} \sim x \log \left( \frac{\log x}{\log g(x)} - 1 \right) \quad \text{as } x \rightarrow \infty.$$

### 2.1.2. The case $f(x) \sim cx^l$ for $l > 1$ and $c > 0$ .

First consider the case of  $f(x) = cx^l + o(x^l)$  for  $c > 0$  and  $g(x) = (x/c)^{1/(l+1)} + o(x^{1/(l+1)})$ . In this case, we see that

$$\begin{aligned} A_{11} &= \int_2^{g(x)} \left( \text{li}(f(t)) + \frac{f(t)}{\log t} \right) dt \\ &= O \left( \int_2^{g(x)} \frac{t^l}{\log t} dt \right) = O \left( \frac{x}{\log x} \right), \\ A_{12} &= x \log \left( \frac{\log x}{\log g(x)} - 1 \right) \\ &= x \log \left( \frac{(l+1) \log x}{\log x - \log c + o(1)} - 1 \right) = x \log l + o(x), \\ A_{13} &= g(x) \text{li}\left(\frac{x}{g(x)}\right) = O \left( \frac{x}{\log x} \right). \end{aligned}$$

Then we have

$$A_1 \sim A_{12} \sim (\log l)x \quad \text{as } x \rightarrow \infty.$$

### 2.1.3. The case of $f(x) \sim cx$ for $c > 1$ .

In this case,  $g(x) = (x/c)^{1/2} + o(x^{1/2})$ . Then we have

$$\begin{aligned} A_{11} &= \int_2^{g(x)} \left( \text{li}(f(t)) + \frac{f(t)}{\log t} \right) dt = \frac{2x}{\log x} + o \left( \frac{x}{\log x} \right), \\ A_{12} &= x \log \left( \frac{\log x}{\log g(x)} - 1 \right) \\ &= x \log \left( 1 + \frac{2 \log c + o(1)}{\log x - \log c + o(1)} \right) \\ &= x \frac{2 \log c}{\log x} + o \left( \frac{x}{\log x} \right), \\ A_{13} &= g(x) \text{li}\left(\frac{x}{g(x)}\right) = \frac{2x}{\log x} + o \left( \frac{x}{\log x} \right). \end{aligned}$$

Thus we get

$$A_1 \sim (2 \log c) \frac{x}{\log x} \quad \text{as } x \rightarrow \infty.$$

### 2.1.4. The case of $f(x) - x \sim cx^\delta$ for $c > 0$ and $0 < \delta < 1$ .

In this case, we see that  $g(x) = x^{1/2} - (c/2)x^{\delta/2} + o(x^{\delta/2})$ . Now we estimate each term as follows.

$$\begin{aligned} A_{12} &= x \log \left( \frac{\log x}{\log g(x)} - 1 \right) \\ &= x \log \left( 1 - \frac{\log \left( 1 - \frac{c}{2}x^{\delta/2} + o(x^{\delta/2}) \right)}{\frac{1}{2} \log x + o(1)} \right) \\ &= x \log \left( 1 + \frac{cx^{\delta/2} + o(x^{\delta/2})}{\log x + o(1)} \right) \\ &= c \frac{x^{\delta/2}}{\log x} + o \left( \frac{x^{\delta/2}}{\log x} \right), \end{aligned}$$

We also have

$$\begin{aligned} &\int_2^{g(x)} \left( \text{li}(f(t)) + \frac{f(t)}{\log t} \right) dt \\ &= \int_2^{g(x)} \left( \text{li}(t + ct^\delta) + \frac{t + ct^\delta}{\log t} \right) dt \\ &\quad + \int_2^{g(x)} \left( \text{li}(o(t^\delta)) + \frac{o(t^\delta)}{\log t} \right) dt \\ &= [t \text{li}(t + ct^\delta)]_2^{g(x)} + \int_2^{g(x)} \left( -\frac{t + c\delta t^\delta}{\log t + O(t^{\delta-1})} \right. \\ &\quad \left. + \frac{t + ct^\delta}{\log t} \right) dt + o \left( \frac{x^{\delta/2}}{\log x} \right) \\ &= g(x) \text{li}(g(x) + cg(x)^\delta) + c(1 - \delta) \int_2^{g(x)} \frac{t^\delta}{\log t} dt \\ &\quad + \int_2^{g(x)} O \left( \frac{t^\delta}{(\log t)^2} \right) dt + o \left( \frac{x^{\delta/2}}{\log x} \right) \\ &= g(x) \text{li}(f(g(x)) + o(g(x)^\delta)) \\ &\quad + 2c \frac{1 - \delta}{1 + \delta} \frac{x^{\delta/2}}{\log x} + o \left( \frac{x^{\delta/2}}{\log x} \right). \end{aligned}$$

Since

$$g(x) \text{li}(f(g(x)) + o(g(x)^\delta)) - A_{13} = o \left( \frac{x^{\delta/2}}{\log x} \right),$$

we obtain

$$A_1 \sim c \frac{3 - \delta}{1 + \delta} \frac{x^{(\delta+1)/2}}{\log x} \quad \text{as } x \rightarrow \infty.$$

## 2.2. ESTIMATE OF $A_2$ .

### 2.2.1. When RH is not assumed.

Since  $f(x)$  is increasing,  $f'(x)$  takes positive values for  $t > 1$ . And we see that  $R_1(x), R_2(x) \ll x/(\log x)^l$  for any  $l \geq 1$ .

Then we have

$$\begin{aligned} A_2 &= \int_2^{g(x)} O\left(\frac{f'(t)t}{(\log t)^L}\right)dt + \int_2^{x^{1/2}} O\left(\frac{t}{(\log t)^L}\right)dt \\ &\quad + \int_{g(x)}^{x^{1/2}} O\left(\frac{x}{t(\log t)^L(\log x - \log t)}\right)dt \\ &= O\left(\int_2^{g(x)} \frac{f'(t)t}{(\log t)^L} dt\right) + O\left(\frac{x}{(\log x)^2}\right) \\ &= O\left(\frac{x}{(\log g(x))^2}\right). \end{aligned}$$

This means that

$$A_2 = \begin{cases} o(x), & f(x) \gg x^M \text{ for any } M > 0, \\ o\left(\frac{x}{\log x}\right), & f(x) \sim cx^l \text{ for } l \geq 1, c > 0. \end{cases}$$

**2.2.2. When RH holds.**

If the Riemann hypothesis is true, it holds that  $R_1(x), R_2(x) \ll x^{1/2+\epsilon}$  for any  $\epsilon > 0$ . Consider the case of  $f(x) \sim x$ . It is not difficult to see that

$$\begin{aligned} A_2 &= \int_2^{g(x)} O(t^{1/2+\epsilon})dt + x \int_{g(x)}^{x^{1/2}} O(t^{-3/2+\epsilon})dt \\ &\quad + \int_2^{x^{1/2}} O(t^{1/2+\epsilon})dt \\ &= O(x^{3/4+\epsilon}). \end{aligned}$$

**2.3. ESTIMATE OF B.**

**2.3.1. When RH is not assumed.**

Since  $R_1(x), R_2(x) \ll x/(\log x)^L$  for any  $L > 0$ , we have

$$\begin{aligned} B &= \sum_{p < g(x)} O\left(\frac{(\log p)f(p)}{(\log f(p))^L}\right) \\ &\quad + \sum_{g(x) \leq p < x^{1/2}} O\left(\frac{x \log p}{p(\log x - \log p)^L}\right) \\ &\quad + \sum_{p < x^{1/2}} \left(\frac{p}{(\log p)^{L-1}}\right) \\ &= O\left(\frac{f(g(x))}{(\log f(g(x)))^L} \sum_{p < g(x)} \log p\right) \\ &\quad + \frac{x}{(\log x)^L} \sum_{g(x) \leq p < x^{1/2}} O\left(\frac{\log p}{p}\right) \\ &\quad + \sum_{p < x^{1/2}} \left(\frac{p}{(\log p)^{L-1}}\right) \\ &= O\left(\frac{x}{\log(x/g(x))^L}\right) + O\left(\frac{x}{(\log x)^{L-1}}\right) \\ &= O\left(\frac{x}{(\log x)^{L-1}}\right). \end{aligned}$$

**2.3.2. When RH holds.**

If the Riemann hypothesis is true, it holds that  $R_1(x), R_2(x) \ll x^{1/2+\epsilon}$  for any  $\epsilon > 0$ . Consider the case where  $f(x) \sim x$ . Then we have

$$\begin{aligned} B &= \sum_{p < g(x)} O((\log p)f(p)^{1/2+\epsilon}) \\ &\quad + \sum_{g(x) \leq p < x^{1/2}} O\left(\log p \frac{x^{1/2+\epsilon}}{p^{1/2+\epsilon}}\right) \\ &\quad + \sum_{p < x^{1/2}} O((\log p)p^{1/2+\epsilon}) \\ &= O(x^{3/4+\epsilon}). \end{aligned}$$

**2.4. CONCLUDING THE PROOF**

Combining the results in Section 2.1, 2.2 and 2.3, we have

$$\begin{aligned} \psi_{2,f}(t) &\sim \begin{cases} x \log\left(\frac{\log x}{\log g(x)} - 1\right), & (f(x) \gg x^M \text{ for } \forall M > 0), \\ (\log l)x, & (f(x) \sim cx^l \text{ for } l > 1 \text{ and } c > 0), \\ (2 \log c) \frac{x}{(\log x)}, & (f(x) \sim cx \text{ for } c > 1), \\ 2c \frac{3 - \delta}{1 + \delta} \frac{x^{\frac{\delta+1}{2}}}{(\log x)}, & (f(x) - x \sim cx^\delta \\ & \text{for } c > 0 \text{ and } 1/2 < \delta < 1, \\ & \text{and RH holds}), \end{cases} \end{aligned}$$

as  $x \rightarrow \infty$ . Since

$$\pi_{2,f}(x) = \int_2^x \frac{d\psi_{2,f}(t)}{\log t},$$

the desired result follows immediately. □

**REFERENCES**

- [1] D. Boneh and G. Durfee, Cryptanalysis of RSA with private key  $d$  less than  $N^{0.292}$ , *Eurocrypt'99, Lecture Notes in Computer Science*, **1592**(1999), 1–11, Springer.
- [2] A. Decker and P. Moree, Counting RSA-integers, *arXiv.math/0801.1451*.
- [3] G.H. Hardy and E. Littlewood, Some problems of 'partitio numerorum'. III: On the expression of a number as a sum of primes, *Acta Math.* **44** (1923), 1–70.
- [4] G.H. Hardy and E.M. Wright, *An introduction to the theory of numbers*, Fifth edition, Oxford University Press, 1979.
- [5] D. Hensley, The distribution of round numbers, *Proc. London Math. Soc.* (3) **54** (1987), 412–444.

- [6] A. Hildebrand and G. Tenenbaum, On the number of prime factors of an integer, *Duke Math. J.* **56** (1988), 471–501.
- [7] A. Ivic, *The Riemann zeta-function. The theory of the Riemann zeta-function with applications*, A Wiley-Interscience Publication, New York, 1985.
- [8] S. Katzenbeisser, Recent Advances in RSA Cryptography, *Advances in Information Security* **3**, Kluwer Academic Publishers (2001).
- [9] J. Korevaar and H. Riele, Average prime-pair counting formula, *arXiv.math/0902.4352*.
- [10] E. Landau, *Handbuch der Lehre von der Verteilung der Primzahlen*, Vol. 1, Leipzig, 1909.
- [11] R.L. Rivest, A. Shamir and L. Adleman, A method for obtaining digital signatures and public-key cryptosystems, *Comm. ACM* **21** (1978), 120–126.
- [12] L.G. Sathe, On a problem of Hardy and Ramanujan on the distribution of integers having a given number of prime factors, *J. Indian Math. Soc.* **17** (1953), 63–141; **18** (1954), 27–81.
- [13] A. Selberg, Note on a paper by L. G. Sathe, *J. Indian Math. Soc.* **18** (1954), 83–87.
- [14] A. Shamir, RSA for paranoids, *RSA Laboratories CryptoBytes*, **1**, no. 3, 1–4, 1995.
- [15] P.W. Shor, Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer, *SIAM J. Comput.* **26** (1997), 1484–1509.
- [16] B. de Weger, Cryptanalysis of RSA with small prime difference, *Appl. Algebra Eng. Commun. Comput.* **13** (2002), 17–28.

Yasufumi Hashimoto

Institute of Systems, Information Technologies and Nanotechnologies, 7F 2-1-22, Momochihama, Fukuoka 814-0001, JAPAN.

E-mail: hasimoto(at)isit.or.jp