

ON INTEGERS WITH TWO PRIME FACTORS

BENJAMIN JUSTUS

ABSTRACT. Integers with two prime factors occur in the RSA cryptosystem. In this paper, we provide density estimates for such integers occurring in the RSA cryptosystem satisfying various conditions. Cryptographic applications are given as a consequence of the estimates obtained.

1. INTRODUCTION

The implementation of the RSA cryptosystem requires the selection of an integer n of the form $n = p \cdot q$ where the distinct prime factors p, q satisfy certain conditions. Such an integer in the literature is often referred as a RSA integer. We follow this convention in the paper. For certain cryptographic applications, it is important to know

- What is the probability that a randomly selected integer is a RSA integer?

In order to answer the question above adequately, one has to know a priori the specific conditions that are imposed on the prime factors p, q of n . A survey of literature shows that no precise and consistent definitions exist. The specific requirements for the the prime factors p, q of n differ among authors. The inventors of the RSA cryptosystem [1, 2] wrote that the primes factors p, q need to be large and be randomly selected. In [9], it is required to select p, q of approximately equal magnitude. In more applied works [3, 10], the authors require p, q to be of equal bit-length.

If one assumes the fact that a randomly selected integer in the interval $[1, x]$ has the probability $\log^{-1} x$ (a consequence of the prime number theorem) of being a prime and furthermore that the events of selecting prime numbers are independent, then one may guess that the answer to the above question is of the order $\log^{-2} x$. This intuition turns out to be true (see Theorem 2.1 and Theorem 3.1) only if one is willing to impose conditions on the prime factors p, q of n . The specific conditions imposed have to do with how close the prime factors p, q are with respect to each other and the tightness of the interval in which p, q are bounded.

Indeed, our original intention of writing the paper is to investigate the question how the density of RSA integers is related the conditions that are imposed upon the prime factors p, q of n . It turns out that in order to obtain the density estimate in the order of $\log^{-2} x$, it is necessary to impose those conditions on the prime factors p, q as described in section 2 and section 3. An early theorem of Landau [6] shows that, the number of integers $n \leq x$ of the form $n = p \cdot q$ with distinct p and q satisfies as x goes to infinity

Received by the editors September 1, 2009.

Key words and phrases. RSA cryptosystem, RSA integer, density estimate.

$$\pi_2(x) \sim \frac{x \log \log x}{\log x}.$$

In particular this result implies that the probability of a randomly selected integer n in the interval $[1, x]$ being of the form $n = p \cdot q$ (no conditions imposed on p, q) is of the order $\frac{\log \log x}{\log x}$.

The organization of the paper is as follows. In section 2 and section 3, we formulate two analytic notions of RSA integers which allow us to quantify: how close the prime factors of a RSA integer are with respect to each other and what we mean by selecting p, q of the same bit-length. We then count respectively the RSA integers satisfying each of the notions. In section 4, we give applications and thereby answer the question that is set out in the introduction.

For the benefit of the readers, we have included appendices at the end of the paper. The appendices contain some well known results from analytic number theory which are used in the paper.

2. FIRST NOTION

The first notion reflects the idea that a RSA integer $n = p \cdot q$ should have its prime factors p, q close to each other. Thus we say,

Definition 1. A RSA integer $n = p \cdot q$ in the interval $[1, x]$ is called θ -spaced if it satisfies the property: if $p < q$, then $p < q \leq x^\theta p$.

In order to provide the density estimate, we need to count θ -spaced RSA integers. Let us consider the following set

$$\mathcal{S}(x; \theta, c) := \{n = p \cdot q \leq x : p < q \leq x^\theta p, p \leq x^c\}.$$

Note: since p is the smaller of the two prime factor we can always take $c \leq \frac{1}{2}$. The main result of the section is

Theorem 2.1. *Let $0 < \theta < 1$ and $0 < c \leq \frac{1}{2}$ be fixed. Then the following estimates for the cardinality of $\mathcal{S}(x; \theta, c)$ hold:*

$$|\mathcal{S}(x; \theta, c)| = \begin{cases} \frac{1}{2c(\theta+c)} \frac{x^{2c+\theta}}{\log^2 x} + O\left(\frac{x^{2c+\theta}}{\log^3 x}\right), & c \leq \frac{1-\theta}{2}; \\ B \frac{x}{\log x} + O\left(\frac{x}{\log^2 x}\right), & \frac{1-\theta}{2} < c \leq \frac{1}{2}. \end{cases}$$

where $B = B(\theta, c)$ is an explicitly computable nonzero constant that depends only on θ and c .

The above theorem shows that how the density of RSA integers changes according to the set parameters θ, c . It should be noticed, in particular, in order to achieve the density in the order of $\log^{-2} x$, the prime factors p, q need to be close (small θ) to each other.

Proof. We deal with the case $c \leq \frac{1-\theta}{2}$ first. Notice $x^\theta p \leq \frac{x}{p}$ if and only if $p \leq x^{\frac{1-\theta}{2}}$. We have

$$|\mathcal{S}(x; \theta, c)| = \sum_{p \leq x^c} \sum_{p < q \leq x^\theta p} 1.$$

The inner sum is treated by the prime number theorem (see Appendix A). Thus

$$|\mathcal{S}(x; \theta, c)| = x^\theta \sum_{\substack{p \\ p \leq x^c}} \frac{p}{\log x^\theta p} + O\left(x^\theta \sum_{\substack{p \\ p \leq x^c}} \frac{p}{\log^2 x^\theta p}\right)$$

Partial summation gives (see Appendix B):

$$\begin{aligned} &= \frac{1}{2c(\theta + c)} \frac{x^{2c+\theta}}{\log^2 x} + O\left(\frac{x^{2c+\theta}}{\log^3 x}\right) + O\left(\frac{x^\theta}{\log^2 x} \sum_{\substack{p \\ p \leq x^c}} p\right) \\ &= \frac{1}{2c(\theta + c)} \frac{x^{2c+\theta}}{\log^2 x} + O\left(\frac{x^{2c+\theta}}{\log^3 x}\right). \end{aligned}$$

This settles the first case. In the second case $\frac{1-\theta}{2} < c \leq \frac{1}{2}$, we have

$$(2.1) \quad |\mathcal{S}(x; \theta, c)| = \sum_{\substack{p \\ p \leq x^{\frac{1-\theta}{2}}}} \sum_{\substack{q \\ p < q \leq x^\theta p}} + \sum_{\substack{p \\ x^{\frac{1-\theta}{2}} < p \leq x^c}} \sum_{\substack{q \\ p < q \leq \frac{x}{p}}} 1.$$

We can bound the first double sum as follows

$$\sum_{\substack{p \\ p \leq x^{\frac{1-\theta}{2}}}} \sum_{\substack{q \\ p < q \leq x^\theta p}} 1 \ll \sum_{\substack{p \\ p \leq x^{\frac{1-\theta}{2}}}} \pi(x^\theta p) \ll x^\theta \sum_{\substack{p \\ p \leq x^{\frac{1-\theta}{2}}}} \frac{p}{\log x^\theta p} \ll \frac{x}{\log^2 x}.$$

The second double sum in (2.1) is the main term. We have

$$\begin{aligned} &\sum_{\substack{p \\ x^{\frac{1-\theta}{2}} < p \leq x^c}} \sum_{\substack{q \\ p < q \leq \frac{x}{p}}} 1 \\ &= \sum_{\substack{p \\ x^{\frac{1-\theta}{2}} < p \leq x^c}} (\pi(x/p) - \pi(p)) \\ &= x \sum_{\substack{p \\ x^{\frac{1-\theta}{2}} < p \leq x^c}} \frac{1}{p \log x/p} - \sum_{\substack{p \\ x^{\frac{1-\theta}{2}} < p \leq x^c}} \frac{p}{\log p} + O\left(x \sum_{\substack{p \\ x^{\frac{1-\theta}{2}} < p \leq x^c}} \frac{1}{p \log^2 x/p}\right). \end{aligned}$$

We have $\sum \frac{p}{\log p} = O(x \log^{-2} x)$ (Example 2, Appendix B) and the term

$$x \sum_{\substack{p \\ x^{\frac{1-\theta}{2}} < p \leq x^c}} \frac{1}{p \log^2 x/p} \ll \frac{x}{\log^2 x} \sum_{\substack{p \\ x^{\frac{1-\theta}{2}} < p \leq x^c}} \frac{1}{p} \ll \frac{x}{\log^2 x}.$$

Thus, (2.1) becomes

$$(2.2) \quad |\mathcal{S}(x; \theta, c)| = x \sum_{\substack{p \\ x^{\frac{1-\theta}{2}} < p \leq x^c}} \frac{1}{p \log x/p} + O\left(\frac{x}{\log^2 x}\right).$$

We are done for the proof except for the evaluation of the sum $\sum \frac{1}{p \log x/p}$. The evaluation of the sum is technical in nature and the proof of which is given at the end this section. Let us for the moment assume the result: for any $0 < \theta < 1$,

$$\sum_{p \leq x^\theta} \frac{1}{p \log x/p} = \frac{\log \log x}{\log x} + \frac{f(\theta)}{\log x} + O\left(\frac{1}{\log^2 x}\right)$$

where $f(\theta)$ is a strictly increasing function in θ . Using this result, (2.2) becomes

$$|\mathcal{S}(x; \theta, c)| = B(\theta, c) \frac{x}{\log x} + O\left(\frac{x}{\log^2 x}\right).$$

The constant $B(\theta, c)$ is positive and nonzero. The theorem is proved. \square

Proposition 2.1. *Let $0 < \theta < 1$. Then the following estimate holds:*

$$\sum_{p \leq x^\theta} \frac{1}{p \log x/p} = \frac{\log \log x}{\log x} + \left(\log \frac{\theta}{1-\theta} + c_1\right) \frac{1}{\log x} + O\left(\frac{1}{\log^2 x}\right)$$

where c_1 is an absolute constant.

Proof. We have

$$\begin{aligned} \sum_{p \leq x^\theta} \frac{1}{p \log x/p} &= \sum_{p \leq x^\theta} \frac{1}{p \log x \left(1 - \frac{\log p}{\log x}\right)} \\ &= \frac{1}{\log x} \sum_{p \leq x^\theta} \frac{1}{p} + \sum_{m \geq 1} \frac{1}{\log^{m+1} x} \sum_{p \leq x^\theta} \frac{\log^m p}{p} \\ (2.3) \quad &= \sum_{m \geq 1} \frac{1}{\log^{m+1} x} \sum_{p \leq x^\theta} \frac{\log^m p}{p} + \frac{\log \log x + c_1 + \log \theta}{\log x} + O\left(\frac{1}{\log^2 x}\right). \end{aligned}$$

The inner sum over p can be dealt with using the following lemma:

Lemma 2.1. *For a positive integer $m \geq 1$,*

$$\sum_{p \leq z} \frac{\log^m p}{p} = \frac{\log^m z}{m} + O(\log^{m-1} z).$$

Proof. The case $m = 1$ is a standard result. For example, the reader may see [7] for a proof. When $m \geq 2$, one has by partial summation

$$\begin{aligned} &\sum_{p \leq z} \frac{\log^m p}{p} \\ &= \left(\sum_{p \leq z} \frac{1}{p}\right) \log^m z - \int_2^z \left(\sum_{p \leq t} \frac{1}{p}\right) d \log^m t \\ &= (\log \log z + c_1) \log^m z - \int_2^z \log \log t (d \log^m t) - c_1 \int_2^z d \log^m t + O(\log^{m-1} z) \\ &= (\log \log z) \log^m z - \int_2^z \log \log t (d \log^m t) + O(\log^{m-1} z) \end{aligned}$$

Performing intergeration by parts to the integral gives

$$= \frac{\log^m z}{m} + O(\log^{m-1} z).$$

This proves the lemma. □

Thus in view of the lemma, (2.3) becomes

$$\begin{aligned} & \sum_{p \leq x^\theta} \frac{1}{p \log x/p} \\ &= \frac{1}{\log x} \sum_{m \geq 1} \frac{\theta^m}{m} + \frac{\log \log x + c_1 + \log \theta}{\log x} + O\left(\frac{1}{\log^2 x} \sum_{m \geq 1} \theta^{m-1}\right) + O\left(\frac{1}{\log^2 x}\right) \\ &= -\frac{\log(1-\theta)}{\log x} + \frac{\log \log x + c_1 + \log \theta}{\log x} + O\left(\frac{1}{\log^2 x}\right) \\ &= \frac{\log \log x}{\log x} + \left(\log \frac{\theta}{1-\theta} + c_1\right) \frac{1}{\log x} + O\left(\frac{1}{\log^2 x}\right). \end{aligned}$$

This proves the proposition. □

The techniques used in the proof can be adapted to more general settings. We briefly mention that in the case $\theta = 0$, there is a result of Decker and Moree [4] in the same spirit.

Corollary 2.1 (Decker, Moree). *Let $C_r(x)$ denote the number of RSA integers $n = p \cdot q$ such that $p < q < rp$, where $r > 1$ is a fixed real number. Then as x tends to infinity, we have*

$$C_r(x) = 2 \log r \frac{x}{\log^2 x} + O\left(\frac{x}{\log^3 x}\right).$$

3. SECOND NOTION

The second notion reflects the idea that the prime factors of a RSA integer should roughly have the same length. Thus, one is lead to consider the following set

$$\mathcal{B}(x; a, b) := \{n = p \cdot q \leq x : x^a < p < q \leq x^b\}.$$

The set parameters a, b describe how small or large the prime factors p, q of a RSA integer are. The main result here is

Theorem 3.1. *Let a, b be two fixed real numbers such that $a < \frac{1}{2}$ and $a < b \leq 1$. Then the following estimates hold:*

$$|\mathcal{B}(x; a, b)| = \begin{cases} \frac{1}{2b^2} \frac{x^{2b}}{\log^2 x} + O\left(\frac{x^{2b}}{\log^3 x}\right), & b \leq \frac{1}{2}; \\ B \frac{x}{\log x} + O\left(\frac{x}{\log^2 x}\right), & b > \frac{1}{2}. \end{cases}$$

where $B = B(b)$ is an explicitly computable nonzero constant depending only on b .

The theorem basically says that, for a fixed a , the density of RSA integers under the current notion really hinges upon what the value b is. In order to achieve the density in the order of $\log^{-2} x$, b must be less than $\frac{1}{2}$.

Proof. We first consider the case $b \leq \frac{1}{2}$. We have

$$\begin{aligned} |\mathcal{B}(x; a, b)| &= \sum_{\substack{p \\ x^a < p \leq x^b}} \sum_{\substack{q \\ p < q \leq x^b}} 1 \\ &= \pi(x^b) \sum_{\substack{p \\ x^a < p \leq x^b}} 1 - \sum_{\substack{p \\ x^a < p \leq x^b}} \pi(p) \\ &= \pi(x^b)^2 - \sum_{\substack{p \\ x^a < p \leq x^b}} \frac{p}{\log p} + O\left(\frac{x^{2b}}{\log^3 x}\right) \end{aligned}$$

now by the prime number theorem and partial summation, this is equal to

$$= \frac{1}{2b^2} \frac{x^{2b}}{\log^2 x} + O\left(\frac{x^{2b}}{\log^3 x}\right).$$

This settles the first case. In the second case $b > \frac{1}{2}$. If $a + b < 1$, then

$$(3.1) \quad |\mathcal{B}(x; a, b)| = \sum_{\substack{p \\ x^a < p \leq x^{1-b}}} \sum_{\substack{q \\ p < q \leq x^b}} 1 + \sum_{\substack{p \\ x^{1-b} < p \leq x^{1/2}}} \sum_{\substack{q \\ p < q \leq x/p}} 1 = I + II.$$

The term I is at most

$$(3.2) \quad I \ll \pi(x^{1-b})\pi(x^b) \ll \frac{x}{\log^2 x}.$$

And the second term II

$$\begin{aligned} (3.3) \quad II &= \sum_{\substack{p \\ x^{1-b} < p \leq x^{1/2}}} (\pi(x/p) - \pi(p)) \\ &= x \sum_{\substack{p \\ x^{1-b} < p \leq x^{1/2}}} \left(\frac{1}{p \log x/p} - \frac{p}{\log p} \right) + O\left(x \sum_{\substack{p \\ x^{1-b} < p \leq x^{1/2}}} \frac{1}{p \log^2 x/p} \right) \\ &= B \frac{x}{\log x} + O\left(\frac{x}{\log^2 x}\right). \end{aligned}$$

The estimate in the last line comes from Proposition 2.1. Therefore the assertion is true in view of (3.2), (3.3) and (3.1).

In the remaining case $a + b \geq 1$,

$$\begin{aligned} |\mathcal{B}(x; a, b)| &= \sum_{\substack{p \\ x^{1-b} < p \leq x^{1/2}}} \sum_{\substack{q \\ p < q \leq x/p}} 1 \\ &= \sum_{\substack{p \\ x^{1-b} < p \leq x^{1/2}}} (\pi(x/p) - \pi(p)) \\ &= B \frac{x}{\log x} + O\left(\frac{x}{\log^2 x}\right). \end{aligned}$$

□

4. GENERATING RSA INTEGERS

In this section, we will answer the question that is set out in the introduction. Suppose first, we are generating RSA integers by picking the prime factors p, q inside a bounded interval. We then have

Theorem 4.1. *Let positive integers m, n and l be fixed such that $m - 1 < l \leq \frac{n}{2}$. Randomly generate a positive integer N with at most n bits. Then the probability of N being a RSA integer whose prime factors have at least m bits and at most l bits is asymptotic to (as $n \rightarrow \infty$)*

$$P(N) = \frac{1}{(l \log 2)^{2^{2n-2l+1}}}.$$

Proof. The set of RSA integers whose prime factors have at least m bits and at most l bits is the set

$$\mathcal{B}\left(2^n; \frac{m-1}{n}, \frac{l}{n}\right) := \{N = p \cdot q < 2^n : 2^{m-1} < p < q < 2^l\}.$$

$|\mathcal{B}(2^n; \frac{m-1}{n}, \frac{l}{n})|$ can be estimated using Theorem 3.1. This gives

$$\left| \mathcal{B}\left(2^n; \frac{m-1}{n}, \frac{l}{n}\right) \right| = \frac{2^{2l}}{2(l \log 2)^2} + O\left(\frac{2^{2l}}{n^3}\right).$$

The theorem readily follows. □

Definition 2. Let s and t be positive integers. We say that s and t are l bits apart if

$$1 < \frac{s}{t} \text{ (or } \frac{t}{s}) \leq 2^l.$$

We may alternatively generate RSA integers by picking one prime first then selecting the other prime near the first prime. We then have the following result.

Theorem 4.2. *Let positive integers m, n and l be fixed such that $2m + l \leq n$. Randomly generate a positive integer N with at most n bits. Then the probability of N being a RSA integer whose prime factors have at most $m + l$ bits and are at most l bits apart is asymptotic to (as $n \rightarrow \infty$)*

$$P(N) = \frac{1}{(\log 2)^2 (ml + m^2) 2^{n-2m-l}}.$$

Proof. The set of RSA integers less than 2^n and whose prime factors have at most m bits and are at most l bits apart has the cardinality twice the size of the following set:

$$\mathcal{S}\left(2^n; \frac{l}{n}, \frac{m}{n}\right) := \{N = p \cdot q < 2^n : p < q < 2^l p, p < 2^m\}.$$

We invoke Theorem 2.1 for the estimate of $|\mathcal{S}(2^n; \frac{l}{n}, \frac{m}{n})|$. Indeed

$$\left| \mathcal{S}\left(2^n; \frac{l}{n}, \frac{m}{n}\right) \right| = \frac{2^{2m+l}}{2(ml + m^2)(\log 2)^2} + O\left(\frac{2^{2m+l}}{n^3}\right).$$

The theorem readily follows. □

5. ACKNOWLEDGEMENTS

The author wishes to thank referees' careful reading of the manuscript. Particular thanks go to Pieter Moree for showing the author how to reduce the error term in the main theorem to $\log^{-2} x$ from $\log \log x \log^{-1} x$ which is what author had originally.

APPENDIX A. THE PRIME NUMBER THEOREM

One usually denotes by $\pi(x)$ the number of primes not exceeding x . We have the following estimate for $\pi(x)$

Theorem A.1. *As x goes to infinity, we have*

$$\pi(x) = \frac{x}{\log x} + O\left(\frac{x}{\log^2 x}\right).$$

We mention that better error terms exist (see [8]) though the error bound $\frac{x}{\log^2 x}$ is good enough for our applications. In particular, the prime number theorem implies

Corollary A.1. *The number of primes in any interval $(x^a, x^b]$ with $a < b$ is*

$$\sum_{x^a < p \leq x^b} 1 = \frac{x^b}{b \log x} + O\left(\frac{x^b}{\log^2 x}\right).$$

The following estimate is needed in the paper. (See [5] for a proof)

Theorem A.2. *There exists a positive constant c such that for $x \geq 2$, one has*

$$\sum_{p \leq x} \frac{1}{p} = \log \log x + c + O\left(\frac{1}{\log x}\right).$$

APPENDIX B. THE METHOD OF PARTIAL SUMMATION

The method of partial summation is a simple but effective tool for handling arithmetic sums.

Theorem B.1. *Let $\langle a_n \rangle$ be a sequence of complex numbers. Set*

$$A(t) = \sum_{n \leq t} a_n \quad (t > 0).$$

Let $b(t)$ be continuously differentiable function on the interval $[1, x]$. Then we have

$$(B.1) \quad \sum_{1 \leq n \leq x} a_n b(n) = A(x)b(x) - \int_1^x A(t)b'(t)dt.$$

The readers can consult [5] for a proof. We illustrate the method by some examples.

Example 1. *As z goes to infinity,*

$$\sum_{p \leq z} p = \frac{z^2}{2 \log z} + O\left(\frac{z^2}{\log^2 z}\right).$$

Indeed in the setting of Theorem B.1, define $\langle a_n \rangle$ by $a_n = 1, n = p$ and $a_n = 0$ otherwise; $b(t) = t$. In view of the prime number theorem, B.1 gives

$$\sum_{p \leq z} p = z\pi(z) - \int_2^z \frac{t}{\log t} dt = \frac{z^2}{\log z} - \int_2^z \frac{t}{\log t} dt + O\left(\frac{z^2}{\log^2 z}\right).$$

Performing integration by parts on the integral, the estimate follows.

Example 2. Let $\theta \geq 0$ and $c > 0$ be fixed. Then x goes to infinity

$$\sum_{p \leq x^c} \frac{p}{\log x^\theta p} = \frac{x^{2c}}{2c(\theta + c) \log^2 x} + O\left(\frac{x^{2c}}{\log^3 x}\right).$$

Define $\langle a_n \rangle$ by $a_n = n$ when $n = p$ and 0 otherwise; $b(t) = \frac{1}{\log x^\theta t}$.

REFERENCES

- [1] Ronald L. Rivest, Adi Shamir, Leonard Adleman. *A method for obtaining digital signatures and public-key cryptosystems*. Technical Report MIT/LCS/TM-82, MIT, Laboratory for Computer Science, Cambridge, Massachusetts, 1977.
- [2] Ronald L. Rivest, Adi Shamir, Leonard Adleman. *A method for obtaining digital signature and public-key cryptosystem*. Communications of the ACM 21(2), 120-126, 1978.
- [3] A. Menezes, P. Van Oorschot, S. Vanstone. *Handbook of Applied Cryptography*. CRC Press, Boca Raton FL, 1997.
- [4] Andreas Decker, Pieter Moree. *Counting RSA-integers*. Results in mathematics, Volume 52, Number 1-2, 2008.
- [5] Gérald Tenenbaum. *Introduction to analytic and probabilistic number theory*. Cambridge studies in advanced mathematics 46, 1995.
- [6] E. Landau, *Handbuch der Lehre von der Verteilung der Primzahlen*. Chelsea Publishing Co., New York, 1953.
- [7] G.H. Hardy, E.M. Wright. *An introduction to the theory of numbers*, fourth Edition. The Clarendon Press, Oxford, 1968.
- [8] A. Ivić. *The Riemann zeta-function*. John Wiley, New York, 1985.
- [9] R. Crandall, C. Pomerance. *Prime numbers a computational perspective*. Springer-Verlag, 2001.
- [10] B. Schneier. *Applied cryptography: protocols, algorithms, and source code in C*. John Wiley & Sons, New York, 2nd Edition, 1996.

UNIVERSITY OF VLORA, DEPARTMENT OF COMPUTER SCIENCE AND ELECTRICAL ENGINEERING,
VLORA, ALBANIA

E-mail address: bjustus@univlora.edu.al