2-1-2019

Estimating the Density of the Abundant Numbers

Dominic Klyve

Melissa Pidde

Kathryn E. Temple

ESTIMATING THE DENSITY OF THE ABUNDANT NUMBERS

Dominic Klyve
Dept. of Mathematics, Central Washington University, Ellensburg, Washington
Dominic.Klyve@cwu.edu

Melissa Pidde
Dept. of Mathematics, Central Washington University, Ellensburg, Washington

Kathryn E. Temple
Dept. of Mathematics, Central Washington University, Ellensburg, Washington
Kathryn.Temple@cwu.edu

Received: 6/10/18, Accepted: 1/18/19, Published: 2/1/19

Abstract

Mathematicians have been interested in properties of abundant numbers – those which are smaller than the sum of their proper factors – for over 2,000 years. During the last century, one line of research has focused in particular on determining the density of abundant numbers in the integers. Current estimates have brought the upper and lower bounds on this density to within about 10^{-4}, with a value of $K \approx 0.2476$, but more precise values seem difficult to obtain. In this paper, we employ computational data and tools from inferential statistics to get more insight into this value. We also put a lower bound on the quantity of abundants in any interval of size 10^6. Finally, we consider the "time series" nature of our data, and consider the possibility of employing tools from this branch of statistics to more carefully refine our statistical estimates.

1. Introduction and Background

In the first century CE, Nichomachus defined an *abundant number* to be an integer the sum of whose proper factors is greater than the integer itself. The smallest abundant number is 12, with factors $1+2+3+4+6 = 16$; there are 22 such integers less than or equal to 100. It is easy to see that there are infinitely many abundant numbers – every multiple of a perfect or abundant number is itself abundant. This leads us quickly to the question: what is the density of the abundant numbers?

It is not immediately apparent that abundant numbers even have a density. In what follows, we shall use the concept of *natural density*, which roughly asks for the fraction of integers which are members of a given set. More formally, we use the following:

Definition 1. Given a set S of positive integers, let $S(n)$ be the number of integers $x \leq n$ with $x \in S$. We define the *natural density* of S, $\mathbf{d}S$, as:

$$\mathbf{d}S := \lim_{n\to\infty} \frac{S(n)}{n},$$

supposing the limit exists [10]. If the limit does not exist, the set does not have a natural density.

Let $\mathcal{A}$ be the set of abundant numbers, and let $\mathbf{d}\mathcal{A}$ represent their (natural) density. With this definition of density, Davenport [3] showed in 1933 that the density of the abundant numbers exists. Finding lower bounds for $\mathbf{d}\mathcal{A}$ is easy. Since every multiple of 6 is abundant, for example, we have $\mathbf{d}\mathcal{A} \geq 1/6$. Similarly, since 20 is the first abundant not a multiple of 6, we can put a stronger lower bound on the density of abundant numbers using a standard inclusion-exclusion argument, noting that $\mathbf{d}\mathcal{A} > 1/6 + 1/20 - 1/60 = 1/5$ (since multiples of 60 are double-counted by the first two terms). Similarly, stronger lower bounds can be found by considering more abundant numbers, again using standard ideas of inclusion and exclusion; we shall revisit this idea in Section 9.1.

Finding upper bounds on $\mathbf{d}\mathcal{A}$ is quite a bit harder; indeed it is not completely trivial to show that $\mathbf{d}\mathcal{A} < 1$. In 1932, however, Behrend [1] showed that $\mathbf{d}\mathcal{A} < 0.47$. In fact, he showed the slightly stronger statement that:

$$\frac{\mathcal{A}(n)}{n} < 0.47$$

for all integers n. Since that time, the best known bounds on $\mathbf{d}\mathcal{A}$ have steadily improved. The progress of the state of the art is shown in Table 1.

Year	Author	Lower	Upper
1932	Behrend [1]		0.47
1933	Behrend [2]	0.241	0.314
1955	Salié [13]	0.246	
$1971-72$	Wall, et al. [15, 14]	0.246	0.2909
1998	Deléglise [4]	0.2474	0.2480
2010	Kobayashi [6]	0.2476171	0.2476475

Table 1: History of the best-known bounds on the density of abundant numbers.

Our goal in this paper is to use computational and statistical methods to explore some questions concerning the distribution of abundant numbers over various ranges. We shall start with the distribution over a small range starting from 1, and then "zoom out" to larger ranges, pausing to describe what we can learn about the density of abundant numbers from each of these snapshots. Over larger ranges, we will naturally use larger subintervals of integers as our unit of study. Rather than try to improve the best known bounds on $\mathbf{d}\mathcal{A}$ (currently due to Kobayashi, whose work involved hundreds of CPU hours, several clever custom algorithms, and was the primary work of his Ph.D. dissertation and a paper [6, 7]), we use our computational and statistical methods to observe several features of the distribution of abundant numbers in various ranges. We shall also introduce the idea of using techniques from the theory of Time Series Analysis to provide insight into the behavior of number-theoretic functions.

While the final interval in Table 1 provides the best proven bounds on the density of abundant numbers, there are some theoretical reasons to think that the true density is at the lower end of Kobayashi's range. We shall thus use this lower bound throughout the paper as a possible conjectural value, and shall call this value $K = 0.2476171$.

2. Values of $\mathcal{A}(n)/n$ to 10^4

As a warm-up to our bigger study, we first examine the density of abundant numbers up to 10,000. Since the first abundant number is 12, we have $\mathcal{A}(n)/n = 0$ for $n < 12$, and thus it is perhaps not surprising to see that $\mathcal{A}(n)/n$ grows fairly monotonically at the beginning.

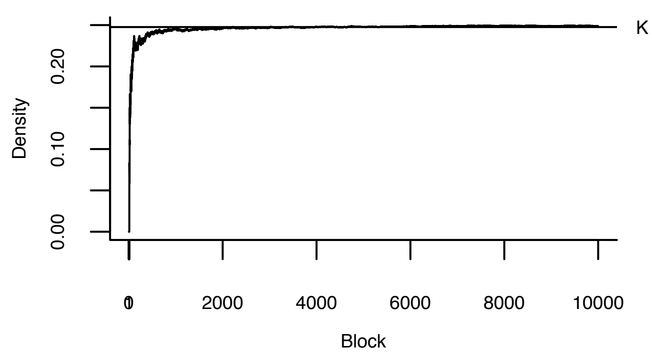

Figure 1: Density of abundant numbers up to 10^4.

Indeed, $\mathcal{A}(n)/n$ is so much smaller for values less than about 10^3 than for larger values this is is difficult to graph all of these on the same axis (see Figure 1). For this reason, we take a closer look at the behavior of $\mathcal{A}(n)/n$ over the region $[10^3, 10^4]$ in Figure 2.

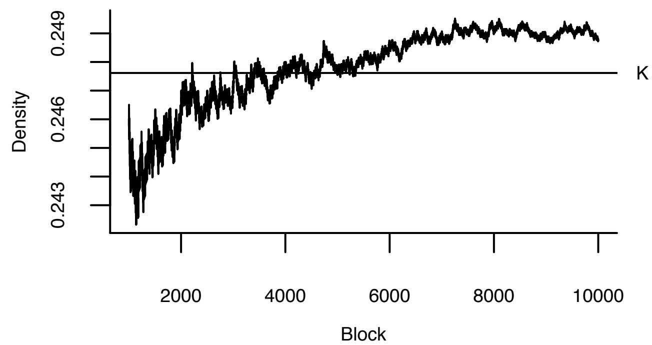

Figure 2: Density of abundant numbers in the range $[10^3, 10^4]$.

We see in this second plot that the density seems to be increasing. Somewhat surprisingly, $\mathcal{A}(n)/n$ passes Kobayashi's upper bound at about $n = 3000$, and does not drop back into his range in the interval $[3000, 10000]$. (Note that Kobayashi's upper and lower bounds on $d\mathcal{A}$ are graphically indistinguishable on the scale of this graph.) Of course, eventually it must do so, and we shall take our first step out in order to try to locate when this might happen.

3. Values of $\mathcal{A}(n)/n$ to 10^7

If we expand our "window" by three orders of magnitude, we find a similar theme on a larger scale. Before we examine the results, however, we establish a few notational conventions we shall follow as we move to larger ranges. In practice, there are computational limits on the number of integers for which we can store the value of $\mathcal{A}(n)$. We therefore collected data about the quantity of abundant numbers in blocks of size 10^6. Expanding on our previous notation, we make the following definition:

Definition 2. We shall use the notation $\mathbf{d}\mathcal{A}_{10^6,n}$ to denote the density of abundant integers in the range $[(n-1)10^6 + 1, n10^6]$.

Thus, $\mathbf{d}\mathcal{A}_{10^6,1}$ is the density of abundant numbers in the first block of one million integers, and $\mathbf{d}\mathcal{A}_{10^6,7}$ is the density of the abundant numbers in the range $[6000001, 7000000]$. A quick computer check gives the values in Table 2.

n	$\mathbf{d}\mathcal{A}_{10^6,n}$
1	0.247545
2	0.247491
3	0.247509
4	0.247766
5	0.247704
6	0.247681
7	0.247713
8	0.247782
9	0.247739
10	0.247807

Table 2: Values of $\mathbf{d}\mathcal{A}_{10^6,n}$ for $1 \leq n \leq 10$. Each gives the proportion of abundant integers in the nth block of size one million.

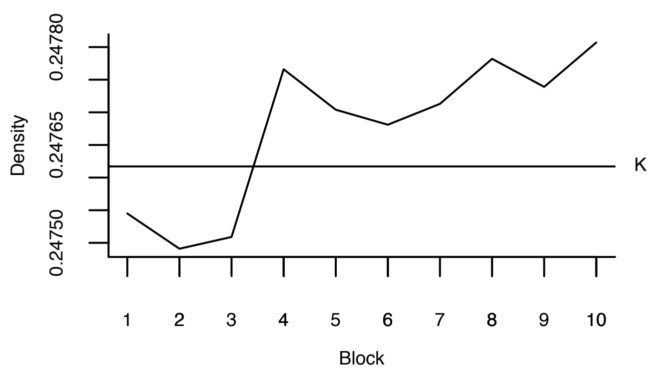

Figure 3: The density of abundant numbers, $\mathbf{d}\mathcal{A}_{10^6,n}$ for $n \in [1, 10]$. The horizontal line is at $K = .2476171$, which is Kobayashi's lower bound. Note that the values are being plotted here for each additional block, and are not cumulative.

As described above, we see in Figure 3 a pattern similar to that seen on the interval $[1, 10^4]$. In the first three blocks of size 10^6, the density of abundant numbers is less than Kobayashi's lower bound, after which they are all greater. Once again, we find that values of $\mathcal{A}(n)/n$ seem to be generally increasing.

It is encouraging that we are able to identify large-scale changes in the density of abundant numbers using these block values. Indeed, from this point on most of the study of abundant numbers in this work will center on these values of $\mathbf{d}\mathcal{A}_{10^6,n}$.

4. Values of $\mathcal{A}(n)/n$ to 10^8

Given that the density of abundant numbers is generally increasing until 10^7, and given that it cannot continue to do so indefinitely, it is reasonable for us to ask when the density levels out or decreases. If we zoom out one more order of magnitude, we find a partial answer to this question; a graph of the block densities $\mathbf{d}\mathcal{A}_{10^6,n}$ for $n \in [1, 100]$ is in Figure 4.

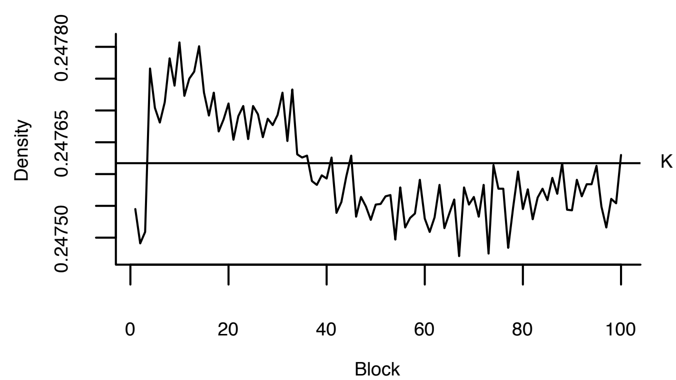

Figure 4: Values of $\mathbf{d}\mathcal{A}_{10^6,n}$ for $n \in [1, 100]$.

The increasing trend, it seems, was short-lived. After about 4×10^7, values of $\mathbf{d}\mathcal{A}_{10^6,n}$ drop below K, and stay below for almost all of these blocks up to 10^8.

A graph of the running averages $\mathcal{A}(n)/n$ is in Figure 5. It shows that, by approximately $80 \cdot 10^6$, $\mathcal{A}(n)/n$ has dropped below K and stays there through the end of this range.

4.1. An Analogy: Counting Primes With $\pi(x)$, $\mathrm{li}(x)$ and $\frac{x}{\log x}$

As we consider the relationship between $\mathcal{A}(n)/n$ and $d\mathcal{A}$, we note that the relationship between the asymptotic limit of a number-theoretic ratio and its finite values can take two general forms. Consider, for example, the relationship between the prime counting function $\pi(x)$, the logarithmic integral $\mathrm{li}(x) := \int_2^x \frac{1}{\log t} dt$, and the function $x/\log(x)$. While it is true that asymptotically both $\pi(x) \sim \mathrm{li}(x)$ and $\pi(x) \sim x/\log(x)$, in the latter case the asymptotic belies a subtle inequality. In fact, for all $x \geq 17, \pi(x) > \frac{x}{\log x}$, and thus $\pi(x) - \frac{x}{\log x} > 0$ for all $x \geq 17$ [12]. On the other hand, Littlewood proved that the sign of $\pi(x) - \mathrm{li}(x)$ changes infinitely often [9] .

Based on the evidence so far, it becomes increasingly tempting to conjecture that $\mathcal{A}(n)$ will oscillate above and below $d\mathcal{A}$ infinitely often. More data might help us

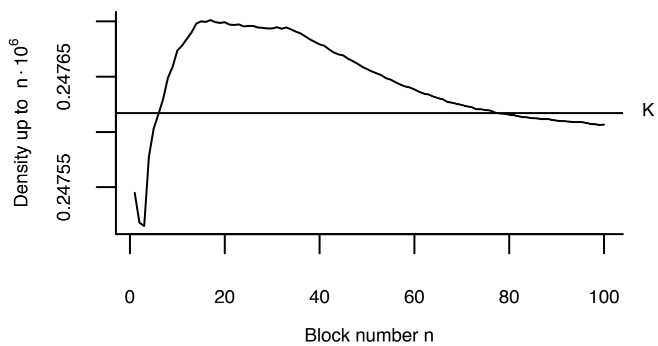

Figure 5: Cumulative density of the abundant numbers up to $n \times 10^6$ for all $n \in [0, 100]$. The horizontal line is at height K.

be more certain, however, and to that end we consider values over a larger scale.

5. Values of $\mathcal{A}(n)/n$ to 10^{11}

We next consider integers up to 10^{11}, giving us 100,000 blocks of size 10^6. At this point, there is enough data that plotting it on a single graph renders the graph unreadable, so instead we include as representative of this data graphs of the first 1000 blocks (integers up to size 10^9), block numbers 9000 - 10,000 (integers between 9×10^9 and 10^{10}) and block numbers 99,000 through 100,000 (integers between 9.9×10^{10} and 10^{11}). These are Figures 6, 7, and 8, respectively.

While there is definitely some initial variability in the densities, this appears to have diminished as the block number increases, and the last two sets of blocks visually appear to be quite similar. There is still substantial block-to-block variation, but overall the pattern does not seem to be changing much, if at all.

Note that because we have an order structure given by the block number (and indirectly by the ordering of the integers), we can think of the data series as being analogous to a time series - a sequence of data points indexed by time. Loosely speaking, statisticians refer to time homogeneity as *stationarity*. We'll discuss stationarity in more detail in Section 7. For now, we simply note that this suggests that it might be useful to apply statistical reasoning and techniques to develop more insight into the behavior of abundant numbers, and specifically into the values $\mathbf{d}\mathcal{A}_{10^6,n}$. In the next section, therefore, we begin to use statistical techniques to study these values.

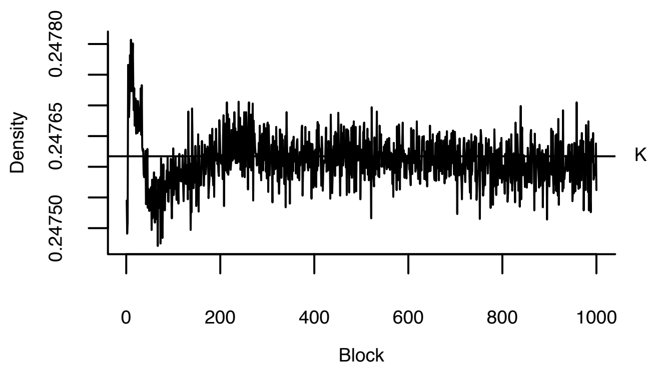

Figure 6: Density of abundant numbers in the first 1000 blocks of size 10^6.

6. A Statistical Approach to Abundant Numbers

Given an integer n, we let $a(n)$ be the function that returns 1 if n is abundant, and 0 otherwise. Intuitively, we might expect $a(n)$ to behave like a random variable taking values from the set $\{0, 1\}$, with $a(n)$ taking the value 1 with probability $\mathbf{d}\mathcal{A}$, and taking the value 0 with probability $1-\mathbf{d}\mathcal{A}$. To that end, let us define just such a random variable, X, as follows:

$$X := \begin{cases} 0 & \text{with probability } 1 - \mathbf{d}\mathcal{A}; \\ 1 & \text{with probability } \mathbf{d}\mathcal{A}. \end{cases}$$

A random variable X which takes two distinct values with given probabilities is a *Bernoulli* variable. It has a mean $\mu_X = \mathbf{d}\mathcal{A}$ and standard deviation $\sigma_X = \sqrt{\mathbf{d}\mathcal{A}(1-\mathbf{d}\mathcal{A})}$. Since we are working under the assumption that X behaves like $a(n)$, we use the midpoint of Kobayashi's bounds on $\mathbf{d}\mathcal{A}$ above to estimate $\mu_X \approx 0.247632$ and $\sigma_X \approx 0.431637$.

Because we are interested in the distribution of values of $\mathbf{d}\mathcal{A}_{10^6,n}$, we also define a new random variable Y to model the proportion of abundant integers in blocks of size 10^6 as

$$Y := \frac{1}{10^6} \sum_{i=1}^{10^6} X_i.$$

We can model the distribution of Y by recalling the Central Limit Theorem.

Theorem 3 (Central Limit Theorem). *Given a set of independent, identically distributed variables $X_1, X_2, \ldots X_n$, whose common distribution has mean μ and*

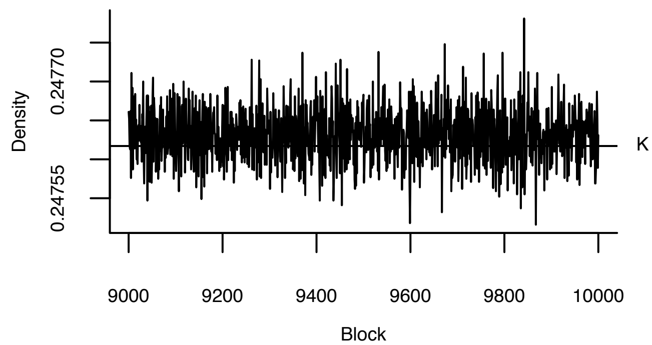

Figure 7: Density of abundant numbers in an interval of 1000 consecutive blocks of size 10^6, starting with block 9000.

finite standard deviation σ, for large n their mean $\bar{X}$ is approximately normally distributed with mean μ and standard deviation $\sigma/\sqrt{n}$; that is, the distribution of $\bar{X}$ can be approximated by the $N(\mu, \sigma/\sqrt{n})$ distribution.

We have previously defined the random variable X, trying to use it as a model for whether a single integer is abundant. Let X_i correspond to the ith integer. Since the variables X_i are indeed independent, identically distributed variables, then Y is approximately normally distributed with mean $\mathbf{d}\mathcal{A}$ and standard deviation $\sqrt{\frac{\mathbf{d}\mathcal{A}(1-\mathbf{d}\mathcal{A})}{10^6}}$.

Using the Central Limit Theorem, we can try to predict the behavior of the values of $\mathbf{d}\mathcal{A}_{10^6,n}$ over different values of n. We have:

$$\mathbf{d}\mathcal{A}_{10^6,n} = \frac{1}{10^6} \sum_{i=(n-1)10^6+1}^{n \cdot 10^6} a(i).$$

If we assume that the $a(i)$ are independent, identically distributed variables, we would expect that Y behaves like $\mathbf{d}\mathcal{A}_{10^6,n}$, or more formally,

Conjecture 4. The values of $\mathbf{d}\mathcal{A}_{10^6,n}$ over all values of n are normally distributed with mean $\mathbf{d}\mathcal{A}$ and standard deviation $\sqrt{\frac{\mathbf{d}\mathcal{A}(1-\mathbf{d}\mathcal{A})}{10^6}}$. Taking the midpoint of Kobayashi's interval as our value for $\mathbf{d}\mathcal{A}$, we expect more precisely that the values of $\mathbf{d}\mathcal{A}_{10^6,n}$ will fall in an $N(0.247632, 0.0004316)$ distribution.

It now remains to check this conjecture against the computational data.

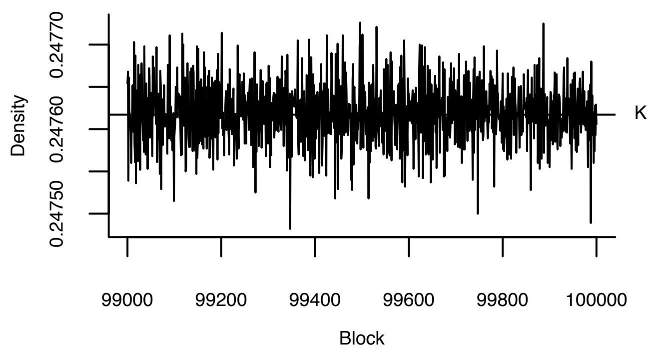

Figure 8: Density of abundant numbers in an interval of 1000 consecutive blocks of size 10^6, starting with block 99000.

6.1. Missing Variability in $\mathbf{d}\mathcal{A}_{10^6,n}$

As expected, the values of $\mathbf{d}\mathcal{A}_{10^6,n}$ are normally distributed (Figure 9). However, the standard deviation is far from what we conjectured above. We expected to find a standard deviation of about 4.32×10^{-4}. Instead, the value is only 3.54×10^{-5} – more than an order of magnitude less. We next need to address the question of why the values of $\mathbf{d}\mathcal{A}_{10^6,n}$ seem to be clustered together so much more than we would expect.

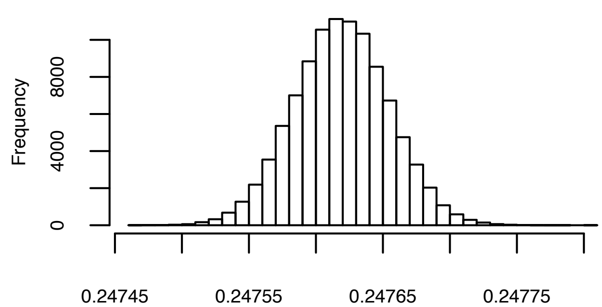

Figure 9: A histogram giving the number of blocks with given density from the values of $\mathbf{d}\mathcal{A}_{10^6,n}$ for all $n \in [1, 100\ 000]$.

In formulating Conjecture 4, we used the hypothesis that the individual values of $a(n)$ were independent. That is, we assumed that whether any particular integer is abundant does not influence whether any other integer is abundant. We first need to examine this assumption more closely.

6.2. Independence of Values of $a(n)$

Recall that our function $a(n)$ is an indicator function on whether an integer n is abundant. Before considering the trickier question of the behavior of the values of $\mathbf{d}\mathcal{A}_{10^6,n}$, let us first look at the independence of $a(n)$. For example, one implication of independence would be that knowing that an integer n is abundant would not change the likelihood that $n+1$ is abundant.

Even without answering the question of independence precisely, we can note several things. The first is that even numbers are considerably more likely to be abundant than odd numbers – this alone suggests that if n is an abundant even integer, $n+1$ is unlikely to be abundant. A related consideration concerns short runs. Since every multiple of 6 is abundant, we see that the abundants occur with more regularity than they would if $a(n)$ truly behaved like a sequence of independent random variables. Therefore, we see that the values of $a(n)$ are not independent, and that we shall therefore need more sophisticated methods to describe the distribution of $\mathbf{d}\mathcal{A}_{10^6,n}$.

6.3. Back to the Distribution of $\mathbf{d}\mathcal{A}_{10^6,n}$

We prove in Section 9 that any block of size 10^6 must necessarily contain at least $237,110$ abundant numbers. This lower bound helps explain the unusually small standard deviation in the values of $\mathbf{d}\mathcal{A}_{10^6,n}$. Since the blocks contain about 247,600 abundant numbers on average, the "random" nature of the distribution will occur entirely in the remaining $10,000$ or so values. That is, we should expect a much smaller standard deviation in the values of $\mathbf{d}\mathcal{A}_{10^6,n}$ than our naïve heuristic suggested.

6.4. Independence of Values of $\mathbf{d}\mathcal{A}_{10^6,n}$

We know that values of $a(n)$ aren't really independent, but what about $\mathbf{d}\mathcal{A}_{10^6,n}$? Because each of these values incorporates information about the abundance of 10^6 different integers, it seems likely that the issues arising from dependence of $a(n)$ will be "smoothed out", letting us make the following conjecture.

Conjecture 5. The values of $\mathbf{d}\mathcal{A}_{10^6,n}$ over all values of n form a sequence of independent, identically distributed variables.

We shall undertake a careful exploration of Conjecture 5 in Section 7.

If Conjecture 5 is true, we can use the first 100,000 blocks described in Section 5 to calculate an approximate 95% confidence interval for the true mean value

$\mathbf{d}\mathcal{A}_{10^6,n}$, and therefore the value of $\mathbf{d}\mathcal{A}$. Over these first 100,000 values, we first find $\bar{x} = 0.24761949309$ and $s = 3.54164 \times 10^{-5}$, and thus our confidence interval becomes

$$0.24761949309 \pm 1.96 * \frac{(3.54164 \times 10^{-5})}{\sqrt{10^5}},$$

or $(0.247619274, 0.247619713)$. Note that this interval lies near the bottom of Kobayashi's interval of $(0.2476171, 0.2476475)$, and thus if our estimate is correct, his lower bound is very near the truth.

However, we must be quite careful here. Some of the graphs presented earlier suggest that there are long-term trends in the values of $\mathbf{d}\mathcal{A}_{10^6,n}$, and that there may be some time series structure to these data.

If the values of $\mathbf{d}\mathcal{A}_{10^6,n}$ are not independent, then the confidence interval calculated above is not valid. We therefore turn next to the field of time series analysis to consider whether the values of $\mathbf{d}\mathcal{A}_{10^6,n}$ are, in fact, independent, and if not to apply some alternate techniques to calculate a confidence interval for $\mathbf{d}\mathcal{A}$.

7. Abundant Densities as Time Series

Since the values $\mathbf{d}\mathcal{A}_{10^6,n}$ have an order structure (they are ordered by the block index n), we can analyze the independence of $\mathbf{d}\mathcal{A}_{10^6,n}$ using tools from time series analysis, where the block index n plays the role of the time variable. A standard tool in time series analysis to look at the order structure of a time series (if any) is the autocorrelation function. The autocorrelation function of a time series Y_t, or ACF, is defined by $\rho(k) = Corr(Y_t, Y_{t+k})$, where $Corr(\cdot, \cdot)$ denotes the correlation between two random variables. The observant reader will note that implicit in this notation is the assumption that the correlation does not depend on the time, t, but only on the distance between the observations, k, which is often referred to as the lag. This will be true provided that the series Y_t is *second-order stationary* or *weakly stationary*, that is, the means of the observations, $E[Y_t]$, variances, $Var(Y_t)$, and covariances, $Cov(Y_t, Y_{t+k})$, are independent of t. For our series $\mathbf{d}\mathcal{A}_{10^6,n}$, the graph suggests that the values from about block number 10,000 to block number 100,000 do, in fact, represent a stationary time series. Since the beginning of the series seems to represent 'burn-in' or initial noisiness not typical of the rest of the series, we will work with the data set starting at block 10,001; this is well past any unusual initial behavior.

Since the data appear to be stationary, we consider the sample ACF, $\hat{\rho}(k)$. If the densities in successive blocks are truly independent, then all of the sample autocorrelations will be near 0. However, $\hat{\rho}(1) = -.169$, so we see that successive blocks are not independent and, in fact, are negatively correlated. Considering edge effects, this seems reasonable: a slightly higher proportion of abundant numbers in

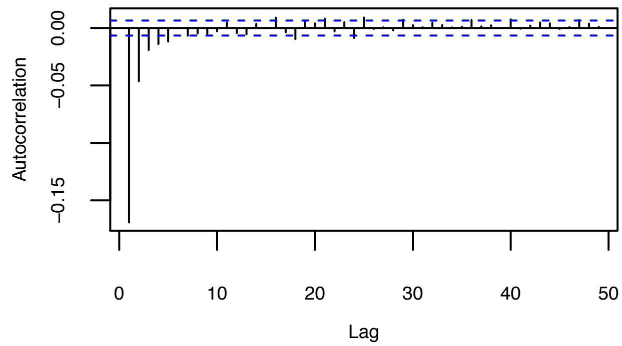

Figure 10: Sample ACF for $\mathbf{d}\mathcal{A}_{10^6,n}$.

one block may mean that it had an extra multiple of 6, 20, etc; the next block then might have a slightly lower proportion of abundant numbers. The second autocorrelation is also negative, but with smaller magnitude, and the magnitude then falls off somewhat quickly (Figure 10).

Unfortunately, what this goes to establish is that the block densities $\mathbf{d}\mathcal{A}_{10^6,n}$ are not independent in n, which means that the conditions for the standard normal-based confidence interval discussed above are not met. However, the problem of giving a confidence interval for the mean of a stationary time series is a standard one in the time series literature. There are various approaches proposed for calculating such a confidence interval; we will implement a bootstrap-based method called the *moving block bootstrap*. This method was chosen because of its relatively mild assumptions and, in particular, the fact that it does not require a particular model for the time series to be specified.

The moving block bootstrap method was originally proposed by Künsch in [8]. (Note that there is no relationship between the general term 'block' in the bootstrap method and our use of it to refer to a set of integers in this paper.) For this method, our original series of 90,000 observations is divided into $(90,000 - b + 1)$ overlapping blocks of block size b. Then $90,000/b$ of these blocks are chosen with replacement and used to form a new series of 90,000 observations. We compute the mean of the new series and then repeat the whole process many times. The resulting means are viewed as a sample from an empirical distribution, and then the bootstrap confidence interval uses the percentiles of this sample to give a confidence interval for the population mean. The table below gives results for several choices of block size from $b = 100$ to $b = 10,000$. In each case, the re-sampling procedure was repeated 10,000 times and the 2.5th and 97.5th percentiles of the resulting empirical distributions were taken as lower and upper confidence limits. Along with the confidence interval itself, the table also gives the width of each confidence interval. Notice that the confidence intervals are consistent with the confidence interval calculated using the standard normal-based techniques, and the narrowest of the intervals is slightly narrower than the classical confidence interval.

Block size	Lower endpoint	Upper endpoint	Width
100	0.24761881	0.24761918	3.7×10^{-7}
500	0.24761867	0.24761930	6.3×10^{-7}
1000	0.24761855	0.24761940	8.5×10^{-7}
5000	0.24761816	0.24761961	1.45×10^{-6}
10,000	0.24761801	0.24761943	1.42×10^{-6}
Classical CI	0.247619274	0.247619713	4.39×10^{-7}

These techniques seem promising, but do require that the time series be stationary (since estimating the mean requires that there be a time-invariant mean to estimate). In our efforts to see whether it is plausible that these conditions hold, we made a final zoom out to integers of size 10^{12}.

8. Values of $\mathcal{A}(n)/n$ to 10^{12}

In an effort to see whether this behavior continues, we then ran an additional 900,000 blocks, giving us 10^6 blocks each of size 10^6 (representing abundant integers out to size 10^{12}). With this much data, however, it turns out that the noisiness between blocks is actually hiding some of the underlying behavior of the data. This is more obvious if we consider somewhat larger blocks. If we take these 10^{12} integers but use blocks of size 10^9 rather than 10^6 (giving us 1000 data points rather than 10^6), we see that our conjecture that the series was a stationary time series is, unfortunately, not the case.

Figure 11 shows some very interesting behavior; the averages are below Kobayashi's lower bound for a sizable range approximately between blocks 200 and 400, corresponding to integers between 2×10^{11} and 4×10^{11}. Beyond this there appears to be a very slow upward trend through the range calculated. We see that the range up to block 100 at this scale (representing the first 100,000 blocks at the previous scale) do appear to be stationary and fairly stable, but that the behavior changes noticeably shortly thereafter.

Unfortunately, this means that further analysis based on looking at the block densities as a stationary time series is not possible, since there is a noticeable slow upward trend. The fact that there seem to be oscillations leads to the following conjecture:

Conjecture 6. The values of $\frac{\mathcal{A}(n)}{n} - d\mathcal{A}$ will change sign infinitely often.

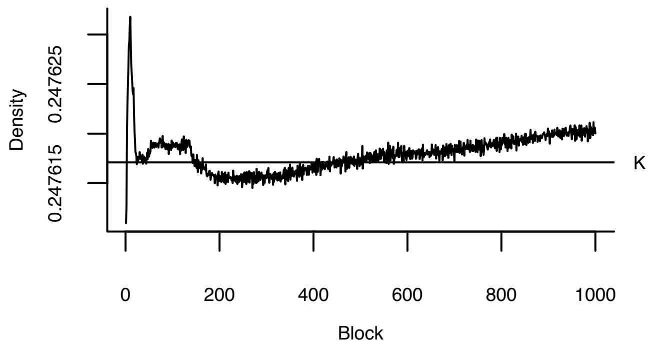

Figure 11: Density of abundant numbers using a block size of 10^9, over the first 1000 blocks. The horizontal line is Kobayashi's lower bound.

Despite the fact that our statistical efforts have not been successful in determining new truths about the behavior of $\frac{\mathcal{A}(n)}{n}$ at infinity, we have found some interesting results concerning bounds on the density of abundant numbers in short intervals, a question which it seems has not been considered before. We shall turn to this in the in final section of this paper.

9. Bounds on the Density of Abundant Numbers in Short Intervals

Having found both some promising leads and some disappointing dead ends in our statistical analysis, we conclude this paper with a pair of theorems which determine upper and lower limits on the number of abundants in any consecutive set of 10^6 integers – that is, we explore what limitations exist on the values of $\mathbf{d}\mathcal{A}_{10^6,n}$.

9.1. A Lower Bound on $\mathbf{d}\mathcal{A}_{10^6,n}$

In Section 1, we included a short inclusion-exclusion argument that the density of abundant numbers is at least $1/5$. We can use a more precise version of this reasoning to put an absolute lower bound on the quantity of abundant numbers in any block of 10^6 consecutive integers. We will use the following notation:

Definition 7. Given a set of integers X, let $f_X(d)$ the number of elements of X divisible by d. If there is no confusion about the set in question, we will shorten this notation simply to $f(d)$.

Now let X be any block of 10^6 consecutive integers. Then $f(6) = f_X(6)$, the number of integers in the set divisible by 6, satisfies $\lfloor 10^6/6 \rfloor \leq f(6) \leq \lceil 10^6/6 \rceil$

(where $\lfloor \cdot \rfloor$ and $\lceil \cdot \rceil$ represent the floor and ceiling functions, respectively), with $f(6)$ taking the higher or lower of these values depending on the value of the smallest element of $X \pmod 6$. Since each multiple of 6 is abundant, we can say that there are at least 166666 abundant numbers in any such block. Furthermore, $f(20)$ satisfies $\lfloor 10^6/20 \rfloor \leq f(20) \leq \lceil 10^6/20 \rceil$ and so on. Using this same idea and our inclusion-exclusion argument from Section 1, we find that $\mathbf{d}\mathcal{A}_X$, the proportion of integers in X which are abundant, can be bounded as

$$\begin{aligned}\mathbf{d}\mathcal{A}_X &\geq \frac{1}{10^6}(f_X(6) + f_X(20) - f_X(60)) \\ &\geq \frac{1}{10^6}(\lfloor 10^6/6 \rfloor + \lfloor 10^6/20 \rfloor - \lceil 10^6/60 \rceil) = .199999.\end{aligned}$$

This lower bound is slightly less than the bound of 1/5 given above. The difference reflects the fact that we are now considering the density of abundant numbers in intervals, and thus each term in the inclusion-exclusion argument has the potential to contribute errors due to rounding. In the following we shall work to balance the benefits of including more terms in the argument with the fact that each additional term carries this potential for rounding error.

If we are going to use more values for our inclusion-exclusion argument (beyond 6 and 20), which values should we choose? The natural choice is the set of *primitive nondeficient* numbers. The set of nondeficient numbers is slightly larger than the set of abundant numbers; it includes both abundant numbers and perfect numbers – those integers precisely equal to the sum of their proper factors. A nondeficient number is *primitive* if it is not a multiple of another nondeficient number.

More generally, we can find explicit bounds on the number of abundant integers in an interval using the Bonferroni inequalities. We shall begin with the general setting for the Bonferroni inequalities as given in [11, Theorem 6.8].

Theorem 8 (Bonferroni inequalities). *Let X be a nonempty, finite set of N objects, and let $P_1, \ldots, P_r$ be properties that elements of X may have. For each subset $I \subset \{1, 2, \ldots, r\}$, let $N(I)$ denote the number of elements of X that have each of the properties indexed by the elements of I. Let N_0 denote the number of elements of X with none of these properties. Then if m is a nonnegative even integer,*

$$N_0 \leq \sum_{k=0}^{m} (-1)^k \sum_{\substack{I \subset \{1,2,\ldots,r\} \\ |I|=k}} N(I),$$

while if m is a nonnegative odd integer,

$$N_0 \geq \sum_{k=0}^{m} (-1)^k \sum_{\substack{I \subset \{1,2,\ldots,r\} \\ |I|=k}} N(I).$$

(Here we take $N(\emptyset)$ to be $|X|$, the cardinality of X.)

We shall take the set X in this theorem to be our set above, that is, a set of 10^6 consecutive integers. We shall let P_j be the property that an integer is divisible by j, so $N(\{n_1, n_2, \ldots, n_k\}) = f(n)$, where n is the least common multiple of $\{n_1, n_2, \ldots, n_k\}$. Then N_0 will be the number of elements of X not divisible by any of $n_1, n_2, \ldots, n_r$.

The value N_0 bears a useful relationship with the number of nondeficient integers (hereafter, nondeficients) in X, which we will denote by $\mathcal{A}_X$. In particular, for any set of nondeficient integers I, the union of the nondeficients in X and those integers counted by N_0 contains all of X, so we have $|X| \leq \mathcal{A}_X + N_0$, or

$$\mathcal{A}_X \geq |X| - N_0. \tag{1}$$

We shall combine these ideas in Theorem 10 to develop a method to find better explicit lower bounds on the density of nondeficients in intervals.

We will first find it useful to have one more definition:

Definition 9. Given a set of integers I, we shall denote by $\mathrm{lcm}(I)$ the least common multiple of all elements of I.

Now applying the Bonferroni inequalities to our setting, we have the following:

Theorem 10. *Let X be a set of n consecutive integers and let d be any positive integer, and recall that $f(k)$ represents the number of elements of X divisible by k. As above, let $\mathcal{A}_X$ denote the number of nondeficient integers in X. Finally, let A_d represent the set of the first d primitive nondeficient integers. Then if m is a nonnegative even integer, we have*

$$\mathcal{A}_X \geq \sum_{\substack{I \subset A_d \\ 0 < |I| \leq m}} (-1)^{|I|+1} f(\mathrm{lcm}(I)),$$

Proof. From Theorem 8, we know that for nonnegative even integers m, we have

$$N_0 \leq \sum_{k=0}^{m} (-1)^k \sum_{\substack{I \subset A_d \\ |I| = k}} N(I).$$

Thus, by (1), the quantity of nondeficient integers in X, $\mathcal{A}_X$, satisfies

$$\mathcal{A}_X \geq |X| - \sum_{k=0}^{m} (-1)^k \sum_{\substack{I \subset A_d \\ |I| = k}} N(I).$$

Note that in our notation, we have that $N(I) = f(\mathrm{lcm}(I))$, and thus we can write this as

$$\mathcal{A}_X \geq |X| - \sum_{k=0}^{m} (-1)^k \sum_{\substack{I \subset A_d \\ |I| = k}} f(\mathrm{lcm}(I)).$$

The double sum above is taken over all subsets of A_d of size at most m. We can thus rewrite the double sum directly, and find

$$\mathcal{A}_X \geq |X| - \sum_{\substack{I \subset A_d \\ |I| \leq m}} (-1)^{|I|} f(\mathrm{lcm}(I)).$$

Finally, note that by the definition of $f_X(n)$, when I is the empty set, we have $(-1)^{|I|} \cdot f(\mathrm{lcm}(I)) = |X|$, and so we can rewrite the above as

$$\mathcal{A}_X \geq - \sum_{\substack{I \subset A_d \\ 0<|I| \leq m}} (-1)^{|I|} f(\mathrm{lcm}(I)) = \sum_{\substack{I \subset A_d \\ 0<|I| \leq m}} (-1)^{|I|+1} f(\mathrm{lcm}(I)),$$

proving the theorem. □

We can use Theorem 10 to find lower bounds on $\mathcal{A}_X$, the number of nondeficients in a set X, by choosing values of m and d. In practice, we will find it useful to replace each of the summands in Theorem 10 with an easily computable lower bound. We shall do this using the following definition.

Definition 11. Given a set of integers I, let $g(I)$ be defined as

$$g(I) := \begin{cases} \lfloor \frac{10^6}{\mathrm{lcm}(I)} \rfloor & \text{if } |I| \text{ is odd;} \\ \lceil \frac{10^6}{\mathrm{lcm}(I)} \rceil & \text{if } |I| \text{ is even.} \end{cases}$$

Note that we now have, for each of the summands in Theorem 10,

$$(-1)^{|I|+1} f(\mathrm{lcm}(I)) > (-1)^{|I|+1} g(I).$$

We therefore have the following.

Theorem 12. *Given X, m, and d as in Theorem 10, we can put a lower bound on the quantity of nondeficients in X, $\mathcal{A}_X$ as*

$$\mathcal{A}_X \geq \sum_{\substack{I \subset A_d \\ 0<|I| \leq m}} (-1)^{|I|+1} g(I).$$

Returning to the particular context in which we are working, we apply this theorem to blocks of size 10^6 and write:

Theorem 13. *Let A_d be the set of the first d primitive nondeficient integers. Then for any nonnegative even integer $m \leq d$ and any $n \geq 0$, we have that the proportion of nondeficient integers in any interval of length 10^6 is at least*

$$\frac{1}{10^6} \sum_{\substack{I \subset A_d \\ 0<|I| \leq m}} (-1)^{|I|+1} g(I).$$

Proof. The theorem follows by applying Theorem 12, setting X to be any set of 10^6 consecutive integers. $\square$

The reader should confirm that choosing $d = 2$, $m = 2$ in Proposition 13 would give the result that the proportion of nondeficient integers in any interval of length 10^6 is at least $\frac{1}{10^6}\left(\lfloor 10^6/6\rfloor + \lfloor 10^6/20\rfloor - \lceil 10^6/60\rceil\right) = 0.199999$.

It remains for us to find the best lower bound by finding the best choices of d and m. In fact, for any given d, there will an "optimal" value of m – that is, a choice of m which gives the best lower bound. The basic idea here is that increasing m gives us more terms in our inclusion-exclusion calculation (and thus a better bound), but includes more rounding errors (and thus a worse bound). In practice, finding a general function for predicting the best value of m for a given d in these cases is difficult or impossible, and instead we computationally check all possible values of m to find the best choice.

In Table 3, for each choice of d, the optimal value of $m \le d$ is shown, together with the resulting bound on the proportion of nondeficient integers in any interval of length 10^6. The bolded row is the optimal lower bound using this technique for those values of d tested (and likely for all d).

d	m	Bound
2	2	0.199999
3	2	0.216665
4	4	0.223803
5	4	0.228882
6	6	0.232969
7	6	0.234347
8	6	0.235469
9	6	0.236277
10	6	0.236765
11	6	0.237024
12	**4**	**0.237111**
13	4	0.236887
14	4	0.236527
15	4	0.236103
16	4	0.235535
17	4	0.235077
18	4	0.234633
19	4	0.234025
20	4	0.233248

Table 3: Best choices of m and corresponding best lower bounds on $\mathbf{d}\mathcal{A}_{10^6,n}$ from Proposition 13 for small values of d.

By choosing $d = 12$ and $m = 4$ in Proposition 13, we can show that 0.237111 is a lower bound of the proportion of nondeficient integers in any interval of length 10^6. (Also note that this seems to be the best possible lower bound using inclusion-exclusion techniques alone – stronger results may be possible, but would require other techniques.) The work and computations above serve to prove the following:

Proposition 14. *Any consecutive sequence of* 10^6 *integers contains at least* 237111 *nondeficient numbers.*

The count of nondeficient numbers, in practice, is very close to the count of abundant numbers (since perfect numbers are rare). It follows from the Euclid-Euler theorem on even perfect numbers that no block of size 10^6 (with smallest value at least 8128) can contain more than such value. (From Table 2, in the first 10^6 integers there are 247545 abundant numbers, so any block with smallest value below 8128 must contain at least 239417 abundant integers.) We can thus also state the following:

Corollary 15. *Assuming the non-existence of odd perfect numbers, any consecutive sequence of* 10^6 *integers contains at least* 237110 *abundant numbers.*

9.2. An Upper Bound on $\mathbf{d}\mathcal{A}_{10^6,n}$

It is not hard to show that the only upper bound on $\mathbf{d}\mathcal{A}_{10^6,n}$ is 10^6, that is, there are 10^6 consecutive abundant numbers. We could, in theory, even find these numbers by using the Chinese Remainder Theorem and solving appropriate modular equations using appropriate "small" abundant numbers.

In fact, Erdős proved the stronger result in 1935 that the longest sequence of consecutive abundant numbers up to n is of the order $\log\log\log n$ [5]. We should thus expect a consecutive sequence of 10^6 abundant numbers somewhere in the range of $\exp(\exp(\exp(10^6)))$!

10. Conclusions

Although the hypotheses required for proper inference using the statistical techniques described in this paper were not satisfied, the fact that all of the estimates made and intervals calculated fell near the bottom of Kobayashi's range suggest that in fact his lower bound may be near the true value of $\mathbf{d}\mathcal{A}$. In fact, we feel comfortable with the conjecture that $0.24761 < \mathbf{d}\mathcal{A} < 0.24762$, which, if true, would fix the value of $\mathbf{d}\mathcal{A}$ to one more decimal digit.

The techniques used in Theorem 13 can be used to put lower bounds on the number of abundant integers in intervals of any length. We note, for example, that taking $d = 22$ and $m = 8$ in the theorem gives that any consecutive sequence of 10^9 integers contains at least $240, 770, 557$ nondeficient numbers.

We also note that these statistical techniques have the potential to be used to address other number-theoretic questions. In particular, we would like to see Time Series Analysis applied to help gain insight into other open problems.

Acknowledgements. The authors would like to thank the referee for careful reading and valuable suggestions which improved the paper.

References

[1] F. Behrend, Über numeri abundantes, *Sitzungsber. Akad. Berlin* (1932), 322–328.

[2] F. Behrend, Über numeri abundantes, ii, *Sitzungsber. Akad. Berlin* (1933), 280–293.

[3] H. Davenport, Über numeri abundantes, *Sitzungsber. Akad. Berlin* (1933), 830–837.

[4] M. Deléglise, Bounds for the density of abundant integers, *Exp. Math.* **7**, no. 2 (1998), 137–143.

[5] P. Erdős, Note on consecutive abundant numbers, *J. Lond. Math. Soc.* **S1-10**, no. 1 (1935), 128.

[6] M. Kobayashi, On the Density of Abundant Numbers, Ph.D. thesis, Dartmouth College, 2010.

[7] M. Kobayashi, A new series for the density of abundant numbers, *Int. J. Number Theory* **10**, no. 1 (2014), 73–84.

[8] H R. Künsch, The jackknife and the bootstrap for general stationary observations, *Ann. Statist.* **17** no. 3 (1989), 1217–1241.

[9] J. E. Littlewood, Sur la distribution des nombres premiers, *C. R. Math. Acad. Sci. Paris* **158** (1914), 1869–1872.

[10] I. Niven, The asymptotic density of sequences, *Bull. Amer. Math. Soc.* **57**, no. 6 (1951), 420–434.

[11] P. Pollack, *Not Always Buried Deep: A Second Course in Elementary Number Theory*, American Mathematical Society, Providence, RI, 2009.

[12] J. B. Rosser and L. Schoenfeld, Approximate formulas for some functions of prime numbers, *Illinois J. Math.* **6** (1962), 64–94.

[13] Hans Salié, Über die Dichte abundanter Zahlen, *Math. Nachr.* **14**, no. 1 (1955), 39–46.

[14] C. R Wall, P. L. Crews, and D. B. Johnson, Density bounds for the sum of divisors function, *Math. Comp.* **26**, no. 119 (1972), 773–777.

[15] C. R. Wall, Topics Related to the Sum of Unitary Divisors of an Integer, Ph.D. thesis, University of Tennessee, 1970.