DENSITY AND SUBSTANCE: AN INVESTIGATION INTO THE SIZE OF INTEGER SUBSETS

ALEX RICE

ABSTRACT. This note was prepared as a final project for Professor Pete L. Clark's elementary number theory course in Spring 2007, the author's third year of undergraduate study at the University of Georgia, and is intended for an undergraduate audience.

When encountering an infinite subset of the natural numbers $\mathbb{N} = \{1, 2, 3, \dots\}$, many questions arise in relation to its frequency or "size" in $\mathbb{N}$. For example, given a truly random natural number (whatever that means), what is the probability (we use this term loosely) that it is prime or square? Many of us are familiar with the divergence of the harmonic series

$$\sum_{n=1}^{\infty} \frac{1}{n},$$

but what can we conclude about the infinite reciprocal summations of various subsets of $\mathbb{N}$? Better yet, what can we conclude from preexisting knowledge of these sums? Here we will offer some more precise definitions and prove some results that will shed light on these questions and more.

For a finite set A, we let $|A|$ denote the cardinality, i.e. number of elements, of A, and we begin with our first nontrivial notion of a subset's size in $\mathbb{N}$.

Definition. For $S \subseteq \mathbb{N}$, we define the *density* of S, which we denote by $\delta(S)$, by

$$\delta(S) = \lim_{N \to \infty} \frac{|S \cap [1, N]|}{N},$$

provided this limit exists.

Example. If $k \in \mathbb{N}$ and $S_k = \{n \in \mathbb{N} : k \mid n\}$, then clearly $\delta(S_k) = 1/k$.

From this definition and example we see that in a loose sense, the density of $S \subseteq \mathbb{N}$ is the probability that a random natural number lies in S. Looking at the numerator $|S \cap [1, N]|$, we are reminded that in the case of the primes, which we denote by $\mathcal{P}$, this is the quantity known as $\pi(N)$, an object of intense study for many years. The famous prime number theorem states that $\pi(N)$ is asymptotic to $N/\log N$, meaning

$$\lim_{N \to \infty} \frac{\pi(N)}{N/\log N} = \lim_{N \to \infty} \frac{\pi(N)}{N} \log N = 1.$$

Since $\log N$ tends to infinity, we must have

$$\lim_{N \to \infty} \frac{\pi(N)}{N} = \delta(\mathcal{P}) = 0.$$

This result may seem counterintuitive, since the event that a random natural number is prime, while unlikely, is certainly not impossible. However, one must consider the impracticality of choose a truly random natural number out of ALL natural numbers, and we recall that in such infinite contexts, events can attain a probability of 0 despite a possibility of occurrence. Such is the case with the primes, and here we have shown in a precise sense that the portion of positive integers that the primes comprise is in fact 0.

A question we have yet to address is one that is ubiquitous when defining quantities as limits. Does the expression $|S \cap [1, N]|/N$ necessarily have a true limit? Equivalently, does every subset $S \subseteq \mathbb{N}$ have a precise density? The answer is a resounding no.

Example. Let $S = \{n \in \mathbb{N} : \text{ the leftmost digit of } n \text{ is } 1\}$.

Notice that for all $N = 2 \cdot 10^m$, $m \in \mathbb{N}$, we have $|S \cap [1, N]|/N > 1/2$,

yet for all $N = 10^m - 1$, $m \in \mathbb{N}$, we have $|S \cap [1, N]|/N = 1/9$.

This is more than enough to conclude that in this case, the limit defining $\delta(S)$ does not exist, so we resort, as we often do, to more reliable notions of the limit.

Definition. For $S \subseteq \mathbb{N}$, we define the *upper density* of S, denoted by $\overline{\delta}(S)$, and the *lower density* of S, denoted by $\underline{\delta}(S)$, by

$$\overline{\delta}(S) = \limsup_{N\to\infty} \frac{|S \cap [1, N]|}{N}, \quad \underline{\delta}(S) = \liminf_{N\to\infty} \frac{|S \cap [1, N]|}{N}.$$

Note that the limit inferior and limit superior exist for any bounded sequence, and our sequences are necessarily bounded between 0 and 1, so we are now armed with quantities to describe all subsets of $\mathbb{N}$. We find in the case of our example that $\overline{\delta}(S) = 5/9$, while $\underline{\delta}(S) = 1/9$. In general, we can offer justification that for all $k \in \{1, 2, \ldots, 9\}$, we have $\overline{\delta}(S_k) = 10/9(k + 1)$ and $\underline{\delta}(S_k) = 1/9k$, where $S_k = \{n \in \mathbb{N} : \text{ the leftmost digit of } n \text{ is } k\}$.

As N grows, $|S_k \cap [1, N]|$ achieves its "local maxima" after a run of integers with first digit k, i.e. values of the form $N = (k + 1)10^m$, $m \in \mathbb{N}$. at these values, the final $1/(k + 1)$ of the integers, as well as exactly $1/9$ of the integers between 1 and $10^m - 1$, all have k as the first digit. Hence

$$\overline{\delta}(S_k) = \lim_{m\to\infty} \frac{10^m - 1}{9(k + 1)10^m} + \frac{1}{k + 1} = \frac{1}{9(k + 1)} + \frac{1}{k + 1} = \frac{10}{9(k + 1)}.$$

The lower density follows a similar argument, using values of the form $N = k \cdot 10^m$, at which $|S_k \cap [1, N]|/N$ attains its "local minima".

This example tells us that with regard to both upper and lower densities, natural numbers that begin with lower digits are "more common" than those that begin with higher digits. This principle holds in practice, and has been used to detect financial fraud. In this example, there was a certain rhyme or reason to the upper and lower densities, but allow us to offer a result that will dispel any potential intuition that there need be any relation between the two besides the obvious inequality.

Claim 1. *For every $\alpha, \beta \in \mathbb{R}$ with $0 \leq \alpha \leq \beta \leq 1$, there exists $S \subseteq \mathbb{N}$ with $\underline{\delta}(S) = \alpha$ and $\overline{\delta}(S) = \beta$.*

Proof. For a real number x, we let $\lfloor x \rfloor$ denote the greatest integer les than or equal to x.

If $\alpha = \beta$, then it is easy to insist that S contains $\lfloor N\alpha \rfloor$ elements of $[1, N]$ for all $N \in \mathbb{N}$, in which case S satisfies $\delta(S) = \alpha$.

If $0 < \alpha < \beta < 1$, choose $N_1 \in \mathbb{N}$ with $\lfloor N_1\beta \rfloor \geq \lfloor N_1\alpha \rfloor + 100$, and choose N_2 such that $\lfloor N_1\beta \rfloor = \lfloor N_2\alpha \rfloor$. Then, inductively choose a sequence $\{N_k\}_{k\in\mathbb{N}}$ of natural numbers strictly increasing to infinity such that

$$\lfloor N_{2k}\alpha \rfloor + (N_{2k+1} - N_{2k}) = \lfloor N_{2k+1}\beta \rfloor = \lfloor N_{2k+2}\alpha \rfloor. \tag{1}$$

Construct $S \subseteq \mathbb{N}$ by insisting that S contains $\lfloor N_1\beta \rfloor$ elements of $[1, N_1]$, no elements of $(N_{2k-1}, N_{2k}]$, and every (integer) element of $(N_{2k}, N_{2k+1}]$ for all $k \in \mathbb{N}$.

We see that $|S \cap [1, N]|/N$ attains its maxima at $N = N_{2k+1}$, at which we have

$$|S \cap [1, N_{2k+1}]| = \lfloor N_{2k+1}\beta \rfloor.$$

Similarly, $|S \cap [1, N]|/N$ attains its minima at $N = N_{2k}$, at which we have

$$|S \cap [1, N_{2k}]| = \lfloor N_{2k}\alpha \rfloor,$$

and the result follows.

If $\alpha = 0$ or $\beta = 1$, we can replace the appropriate conditions in (1) with $N_{2k+2} > N_{2k+1}^2$ or $N_{2k+1} > N_{2k}^2$, respectively. □

We now proceed to a different characterization of the size of a subset of $\mathbb{N}$.

Definition. We say that $S \subseteq \mathbb{N}$ is *substantial* if $\displaystyle\sum_{n\in S} \frac{1}{n}$ diverges.

Intuitively, we interpret that substantial subsets are the "large" subsets of $\mathbb{N}$, but recall that in fact $\sum_{p\in\mathcal{P}} \frac{1}{p}$ diverges. We previously showed that the primes have density 0, which in terms of density is as small as it gets. This casts doubt over what potential relationship these two measures of integer subsets could have to each other, basically contradicting any pleasing biconditional claim we could have investigated. However, we have seen that this notion of substance has allowed us to categorize and distinguish between different subsets of the same density, namely 0. Perhaps substance can serve as a refining property, allowing us to be more precise with our characterization of an integer subset's size. A natural question arises: can the testing of reciprocal sums subdivide the group of integer subsets of any particular density, or do all unsubstantial subsets necessarily have density 0?

So glad you asked.

Claim 2. *If $S \subseteq \mathbb{N}$ and $\overline{\delta}(S) > 0$, then S is substantial.*

Proof. Suppose $S \subseteq \mathbb{N}$ and $\overline{\delta}(S) = \delta > 0$. Let $N_0 = 1$ and choose a sequence $\{N_k\}_{k\in\mathbb{N}}$ of natural numbers such that $N_k \geq 4N_{k-1}/\delta$ and $|S \cap [1, N_k)| \geq \delta N_k/2$ for all $k \in \mathbb{N}$.

Then,

$$\sum_{n\in S}\frac{1}{n}=\sum_{k=1}^{\infty}\sum_{n\in A\cap[N_{k-1},N_k)}\frac{1}{n}\geq\sum_{k=1}^{\infty}\Big(|A\cap[1,N_k)|-N_{k-1}\Big)\frac{1}{N_k}$$
$$\geq\sum_{k=1}^{\infty}(\frac{\delta}{2}N_k-\frac{\delta}{4}N_k)\frac{1}{N_k}=\sum_{k=1}^{\infty}\frac{\delta}{4}\to\infty.$$

□

So the result is proven, and we can equivalently state that if $\sum_{n\in S}\frac{1}{n}$ converges, then we must have $\delta(S)=0$. A classic example of such an integer subset is the squares. Recall that $\sum_{n=1}^{\infty}\frac{1}{n^2}=\frac{\pi^2}{6}$. We certainly don't need this result to conclude that the squares have density 0, but we can actually use this infinite sum and some useful identities to find the density of another, less explicit integer subset.

Definition. A natural number n is called *squarefree* if $p^2\nmid n$ for all $p\in\mathcal{P}$.

If we fix S to be the set of squarefree natural numbers, what would we expect for $\delta(S)$? For each individual prime p, the probability that a natural number is divisible by p^2 is $1/p^2$, and an integer's divisibility by two relatively prime integers are independent events. By this loose probabilistic approach, we expect,

$$\delta(S)=\prod_{p\in\mathcal{P}}\Big(1-\frac{1}{p^2}\Big).$$

We can actually evaluate this product by considering its inverse, and recognizing the general term as a geometric series.

$$\prod_{p\in\mathcal{P}}\Big(\frac{1}{1-\frac{1}{p^2}}\Big)=\prod_{p\in\mathcal{P}}\sum_{k=0}^{\infty}\frac{1}{p^{2k}}=\Big(1+\frac{1}{p_1^2}+\frac{1}{p_1^4}+\cdots\Big)\Big(1+\frac{1}{p_2^2}+\frac{1}{p_2^4}+\cdots\Big)\ldots$$

Since we are dealing entirely with absolutely convergent series, which allows us to rearrange terms, and because every square can be written uniquely as a product of even powers of primes, we can conclude

$$\prod_{p\in\mathcal{P}}\sum_{k=0}^{\infty}\frac{1}{p^{2k}}=\sum_{n=1}^{\infty}\frac{1}{n^2}.$$

In fact, we have just derived a special case of the Euler Product Formula, which says that if a function $F:\mathbb{N}\to\mathbb{C}$ is multiplicative, i.e. satisfies $F(ab)=F(a)F(b)$ whenever $\gcd(a,b)=1$, and the series $\sum_{n=1}^{\infty}|F(n)|$ converges, then

$$\sum_{n=1}^{\infty}F(n)=\prod_{p\in\mathcal{P}}\sum_{k=0}^{\infty}F(p^k).$$

Putting these facts together, we have heuristically argued that

$$\delta(S)=\Big(\prod_{p\in\mathcal{P}}\Big(\frac{1}{1-\frac{1}{p^2}}\Big)\Big)^{-1}=\Big(\sum_{n=1}^{\infty}\frac{1}{n^2}\Big)^{-1}=\frac{6}{\pi^2}.$$

We now proceed in a more rigorous and quantitative fashion to verify this heuristic. Recall that for three functions $f, g, h : \mathbb{N} \to [0, \infty)$, we write

$$f(N) = g(N) + O(h(N))$$

if there exists a constant C such that

$$|f(N) - g(N)| \leq Ch(N)$$

for all $N \in \mathbb{N}$.

Claim 3. *If S is the set of all squarefree natural numbers, then*

$$|S \cap [1, N]| = \frac{6}{\pi^2} N + O(\sqrt{N}),$$

and in particular $\delta(S) = \frac{6}{\pi^2}$.

Proof. We begin by noting that the quantity $|S \cap [1, N]|$ can be expressed in terms of a classic arithmetic function, the Möbius function μ, defined by

$$\mu(n) = \begin{cases} (-1)^m & \text{if } n \text{ is a product of } m \text{ distinct primes} \\ 0 & \text{else} \end{cases}.$$

In particular, we see that $|\mu(n)|$ is 1 is n is squarefree and 0 otherwise, so

$$|S \cap [1, N]| = \sum_{n=1}^{N} |\mu(n)|.$$

Also, we see that every natural number can be written uniquely as $n = rs^2$ where r is squarefree, and hence n is squarefree if and only if $s = 1$. It is clear that $d^2 \mid n$ if and only if $d \mid s$, and we now make use of the ubiquitous identity

$$\sum_{d|k} \mu(d) = \begin{cases} 1 & \text{if } k = 1 \\ 0 & \text{else} \end{cases},$$

from which we have

$$|\mu(n)| = \sum_{d|s} \mu(d) = \sum_{d^2|n} \mu(d).$$

We define a function f on $\mathbb{N}$ by

$$f(d) = \begin{cases} \mu(m) & \text{if } d = m^2, m \in \mathbb{N} \\ 0 & \text{else} \end{cases},$$

so

$$|\mu(n)| = \sum_{m^2|n} \mu(m) = \sum_{d|n} f(d),$$

and hence

$$|S \cap [1, N]| = \sum_{n=1}^{N} \sum_{d|n} f(d).$$

To simplify this double sum, we see that since we are evaluating f at each factor of each integer up to N, we are in fact evaluating f at each integer $d \leq N$ a total of $\lfloor N/d \rfloor$ times. Since $N/d - 1 \leq \lfloor N/d \rfloor \leq N/d$, and $|\sum_{d=1}^{N} f(d)| \leq \sqrt{N}$, we have

$$|S \cap [1, N]| = \sum_{n=1}^{N} \sum_{d|n} f(d) = \sum_{d=1}^{N} f(d) \lfloor \frac{N}{d} \rfloor = \sum_{d=1}^{N} \frac{f(d)}{d} N + O(\sqrt{N}).$$

Substituting back the definition of f, we have

$$|S \cap [1, N]| = \sum_{m=1}^{\sqrt{N}} \frac{\mu(m)}{m^2} N + O(\sqrt{N}).$$

We see that

$$|\sum_{m=\sqrt{N}}^{\infty} \frac{\mu(m)}{m^2}| \leq \sum_{m=\sqrt{N}}^{\infty} \frac{1}{m^2} < \int_{\sqrt{N}-1}^{\infty} \frac{1}{u^2} du = \frac{1}{\sqrt{N} - 1},$$

and hence

$$|S \cap [1, N]| = \sum_{m=1}^{\infty} \frac{\mu(m)}{m^2} N + O(\sqrt{N}).$$

Letting $F(m) = \mu(m)/m^2$, we see that F is a multiplicative function with an absolutely convergent series, so by the aforementioned Euler Product Formula, we have

$$\sum_{m=1}^{\infty} F(m) = \prod_{p \in \mathcal{P}} \sum_{k=0}^{\infty} F(p^k) = \prod_{p \in \mathcal{P}} \Big(1 - \frac{1}{p^2}\Big),$$

which is precisely the expected probability that we previously computed to be $6/\pi^2$, and the result follows. □

Our probabilistic intuition is thus rigorously verified, although perhaps our initial suspicion about the frequency of squarefee integers is betrayed. The author, for one, would certainly not have suspected that the prime factorization of well over half all all natural numbers consists entirely of distinct primes raised to the first power. We continue along this path and investigate the density of other interesting sets in number theory.

Claim 4. *If $S = \{x^2 + y^2 > 0 : x, y \in \mathbb{Z}\}$, then $\delta(S) = 0$.*

Proof. For this "proof", we will stick to the slightly non-rigorous probabilistic approach. We recall that a natural number is the sum of two squares if and only if all of its prime factors that are congruent to 3 modulo 4 occur in its factorization with even exponent. With this in mind, we fix a prime p, and we compute the probability that p occurs in the factorization of a natural number n with even exponent.

This calculation involves an inclusion-exclusion process, where we start with the portion of natural numbers divisible by p^0 (all of them), then subtract the portion divisible by p, then add back the portion divisible by p^2, etc., yielding a probability of

$$1 - \frac{1}{p} + \frac{1}{p^2} - \frac{1}{p^3} + \frac{1}{p^4} \cdots = \sum_{k=0}^{\infty} \Big(-\frac{1}{p}\Big)^k = \frac{1}{1 + \frac{1}{p}} = 1 - \frac{1}{p+1}.$$

Again noting that the required conditions for distinct primes are independent events, we have

$$\delta(S) = \prod_{\substack{p\in\mathcal{P}\\ p\equiv 3 \bmod 4}} \Big(1 - \frac{1}{p+1}\Big).$$

Recall that given a sequence $\{a_k\}_{k\in\mathbb{N}} \subset [0,1)$, the product $\prod_{k=1}^{\infty}(1-a_k)$ is positive if and only if the sum $\sum_{k=1}^{\infty} a_k$ converges. However, it is a fact that the reciprocal sum of the primes in any congruence class which admits infinitely many primes is divergent. In particular,

$$\sum_{\substack{p\in\mathcal{P}\\ p\equiv 3 \bmod 4}} \frac{1}{p+1} \geq \frac{1}{2} \sum_{\substack{p\in\mathcal{P}\\ p\equiv 3 \bmod 4}} \frac{1}{p} \to \infty,$$

and the result follows. □

Claim 5. *If $S = \{x^2+y^2+z^2 > 0 : x,y,z \in \mathbb{Z}\}$, then $\delta(S) = 5/6$.*

Proof. Don't blink, you might miss it!

Recall the three squares theorem, which states that a natural number can be written as the sum of three squares unless it is of the form $4^a(8k+7)$ for some $a,k \in \mathbb{Z}$, or equivalently is congruent to $7(4^a)$ modulo $8(4^a)$. As these congruence classes are disjoint, the portion of natural numbers that are of this form reduces to a geometric series, and we have

$$\delta(S) = 1 - \frac{1}{8}\sum_{a=0}^{\infty}\frac{1}{4^a} = 1 - \frac{1}{8}\Big(\frac{1}{1-\frac{1}{4}}\Big) = 1 - \frac{1}{6} = \frac{5}{6}.$$

□

These results yield an interesting summary of the ability to express natural numbers as the sum of squares. Specifically, if $S_k = \{n \in \mathbb{N} : n \text{ can be written as the sum of } k \text{ squares}\}$, then we have

$$\delta(S_1) = 0, \quad \delta(S_2) = 0, \quad \delta(S_3) = \frac{5}{6}, \quad \text{and} \quad \delta(S_4) = 1.$$

The first equality follows from the convergent reciprocal sum of the squares and Claim 2, while the last equality follows from the four squares theorem, which states that $S_4 = \mathbb{N}$.

These results, along with many other facts that we have unearthed, have significant eyebrow-raising potential. We are now armed with tools to characterize subsets of natural numbers in ways well beyond ubiquitous adjectives like nonempty, finite, or infinite. These tools have allowed us to gain a better understanding of both elementary and subtle integer subsets while both confirming and betraying various intuitions, a phenomenon which, in the author's opinion, is a central motivator of mathematics.

Speaking of motivations, we conclude by defining one final notion of the frequency of an integer subset. Here we convey some groundbreaking results and intriguing conjectures concerning the interplay between this new property and the ones that we have already defined.

Definition. For $k \in \mathbb{N}$, a *k-term arithmetic progression* is a set of the form $\{x, x+d, x+2d, \ldots, x+(k-1)d\}$ with $x,d \in \mathbb{N}$. A set $S \subseteq \mathbb{N}$ is said to *contain arbitrarily long arithmetic progressions* if it contains a k-term arithmetic progression for every k.

Theorem (Erdős-Turán Conjecture #1, 1936/ Szemerédi's Theorem, 1975)**.** *If $S \subseteq \mathbb{N}$ and $\overline{\delta}(S) > 0$, then S contains arbitrarily long arithmetic progressions.*

This result tells us that the existence of arbitrarily long arithmetic progressions is similar to the notion of substance, in that it is weaker than any notion of positive density, and hence partitions the sets of density 0. However, phrasing this result in such context does not do it justice, as the proof is nowhere near as elementary. In fact, for decades after Erdős and Turán made this conjecture, it was one of the great unsolved problems in number theory. Endre Szemerédi unlocked the mystery with a combinatorial proof nearly 40 years after the original conjecture.

As momentous as this result is on its own, it raises yet another question about the relationship between characterizations of the size of integer subsets. So far, our strongest notions of a "large" integer subset are the various forms of positive density, followed by two weaker notions, reciprocal sum divergence and arbitrarily long arithmetic progressions, which partition the collection of density 0 sets. How do these two characterizations compare to one another? In what order do these complete our hierarchy? Or, perhaps, are these two properties equivalent?

Intuition suggests that the existence of arbitrarily long arithmetic progressions is far too mysterious, and the proofs associated to it far too sophisticated, for the property to be equivalent to something as elementary as divergent reciprocal sums. In a refreshing change of pace, this intuition holds up, as we can promptly dispel the possibility of equivalence.

Claim 6. *There exists $S \subseteq \mathbb{N}$ such that S contains arbitrarily long arithmetic progressions and S is not substantial.*

Proof. The construction of such a set is admittedly rather arbitrary, as it is not difficult to form increasingly long arithmetic progressions out of sufficiently huge integers, but here we exhibit a cooked up example with a nice sharpness to it.

We define a sequence $\{n_k\}_{k\in\mathbb{N}} \subset \mathbb{N}$ by $n_k = \frac{k(k+1)}{2} + k - 1$. We see that $n_{k+1} - n_k = k + 2$ for each $k \in \mathbb{N}$, and hence for each $n \in \mathbb{N}$, there exists a unique $k \in \mathbb{N}$ and $j \in \{0, 1, \ldots, k+1\}$ such that $n = n_k + j$.

We now let $S = \{a_n\}_{n\in\mathbb{N}}$ where $a_n = a_{n_k+j} = (n_k + 1)^2 + j(2n_k + 1)$. One can check that $a_n > n^2$ for every $n \in \mathbb{N}$, so we have by a simple comparison test that S is not substantial.

However, by construction the pairs of consecutive terms from a_{n_k} through $a_{n_{k+1}-1}$ have a common difference of $2n_k + 1$, and hence form an arithmetic progression of length $k + 2$. Therefore, we can further out in our sequence and find arithmetic progressions as long as we want. Moreover, we know exactly where to find them, which isn't always so easy!

To avoid any confusion, here are the first several terms of our sequence:

$a_1 = 4 \quad a_2 = 7 \quad a_3 = 10 \quad$ (arithmetic progression of length 3)

$a_4 = 25 \quad a_5 = 34 \quad a_6 = 43 \quad a_7 = 52 \quad$ (arithmetic progression of length 4)

$a_8 = 81 \quad a_9 = 98 \quad a_{10} = 115 \quad a_{11} = 132 \quad a_{12} = 149 \quad$ (arithmetic progression of length 5).

Note the increasingly long progressions, and that each term is larger than its corresponding square. □

So we have made some progress in completing our totem pole by explicitly showing that the existence of arbitrarily long arithmetic progressions is not a stronger characterization of a large integer subset than a divergent reciprocal sum, but what about the converse? Erdős and Turán made an additional conjecture addressing this exact inquiry.

Conjecture (Erdős-Turán Conjecture #2)**.** *If $S \subseteq \mathbb{N}$ is substantial, then S contains arbitrarily long arithmetic progressions.*

This question is still very much open. From Szemerédi's Theorem, the mystery remains only in substantial integer subsets of density 0. It just so happens that the most important integer subset in all of number theory fits this description, so it might be a good place to start. This task was quite famously, and quite recently, tackled by Ben Green and Terence Tao.

Theorem (Green-Tao Theorem, 2004)**.** *The primes contain arbitrarily long arithmetic progressions.*

Amidst the thunderous echoes of this theorem, we gain an understanding of the depth and complexity of what originally seemed like such a simple question: given a subset of the natural numbers, how big is it? We began our journey by stating that the primes were so sparse that they represented no positive proportion of the natural numbers, and we conclude by stating that the primes are frequent enough that if you want $1,000,000$ primes equally spaced apart, then they are out there for you to find (good luck with that by the way). We hit some intriguing and surprising turns along the way, and even more excitingly, we see that many questions remain and the journey is far from over.

Acknowledgements

The author would like to express his infinite gratitude for the mathematical as well as personal support provided by Pete Clark, Ted Shifrin, Neil Lyall, Patrick Corn, Malcolm Adams, and all his friends and colleagues in the University of Georgia Mathematics Department, without whom he may still be a journalism major.

References

[1] H. L. Montgomery, I. Niven, H. S. Zuckerman, *An Introduction to the Theory of Numbers*, John Wiley & Sons, Inc., Fifth Edition, 1991.

[2] P. L. Clark, *Arithmetical Functions III: Orders of Magnitude.*

[3] P. L. Clark, *The Primes: Infinitude, Density, and Substance.*

[4] P. L. Clark, *The Prime Number Theorem and the Riemann Hypothesis.*

All Clark notes from Spring 2007 number theory course webpage, http://www.math.uga.edu/ pete/teaching.html.