BINOMIAL PROBABILITY OF PRIME NUMBER OF SUCCESSES

SUNGJIN KIM AND NILOTPAL KANTI SINHA

ABSTRACT. We prove that the unconditional asymptotic formula for the sum of the binomial coefficients $\sum_p \binom{n}{p}$ over prime numbers $p \le n$ holds for almost all n. We also establish an upper estimate of this sum. Then we show that a nontrivial lower estimate would imply a bound for prime gaps $g_n \ll \sqrt{p_n}$, which is stronger than Cramér's bound $g_n \ll \sqrt{p_n}\log p_n$ conditional on the Riemann Hypothesis.

1. INTRODUCTION

The identities on the multisection sum of the binomial coefficients such as $\sum_{k\ge 0}\binom{n}{ak}$ for $a \ge 1$, arise in various areas, such as combinatorics and applied probability. In 1834, Rasmus proved the general identity on summation of binomial coefficients in arithmetic progression,

$$\sum_{k\ge 0}\binom{n}{ak+b} = \frac{1}{a}\sum_{k=1}^{a}\omega^{-bk}(1+\omega^k)^n = \frac{2^n}{a}\sum_{k=1}^{a}\cos^n\frac{k\pi}{a}\cos\frac{(n-2b)k\pi}{a},$$

where $0 \le b < a, n \ge 0$ and $\omega = e^{2\pi i/a}$ is a primitive ath root of unity. This identity expresses a combinatorial sum in terms of a trigonometric sum. In a similar spirit, we can ask about the sum

$$S_n = \sum_{k\in A}\binom{n}{k},$$

where A is a subset of the set of natural numbers. For various subsets A, we can investigate several interesting properties of the binomial sums taken over the elements of the set A. For example, A can be the set of prime numbers or the set of squares or the set of integers which are coprime to n. Then $S_n/2^n$ is the probability that the number of successes in n independent Bernoulli trials; where the probability of success in each trial is $1/2$; is a number which belongs to the set A.

1.1. **Summation over primes.** In this paper, we consider the case where the summation is taken over all prime numbers $p \le n$ so that $S_n = \sum_{p\le n}\binom{n}{p}$. We could not find any reference in the literature about the sum of the prime binomial coefficients, so we believe that this is a new problem. Since the sum of the first n binomial coefficients is 2^n and there are approximately $\frac{n}{\log n}$ primes $\le n$, very simplified heuristics suggest that roughly $\frac{2^n}{\log n}$ of the contribution to S_n must come from primes. To test this initial guess, we computed the ratio $\frac{S_n \log n}{2^n}$ and its value was found to be close to 1 for most values of n. What was unexpected however was that the distribution of $\frac{S_n \log n}{2^n}$ was found to be very close to normal as shown in Section 2. Based on the experimental evidence, Nilotpal Kanti Sinha posted a problem on Mathematics Stack Exchange (MSE) [S] asking for an asymptotic formula for S_n and remarked that the sum must be about $\frac{2^n}{\log n}$. Sungjin Kim posted an answer with a conjectural lower bound and an unconditional upper bound:

$$\frac{2^n}{\log n} \ll S_n \ll \frac{2^n \log\log n}{\log n}.$$

Through further analysis, the upper bound was subsequently improved to $\frac{2^n\sqrt{\log\log n}}{\log n}$ and to $\frac{2^n}{\log n}$, which is the same order of magnitude as the conjectural lower bound. A deeper analysis revealed that evaluating the true asymptotics of S_n is much more difficult as it depends upon the precise knowledge of the distribution of primes in short intervals around the central binomial coefficient and the gap between consecutive primes. This work was made possible from an insightful comment by Qiaochu Yuan, who remarked that S_n would be dominated by contributions from terms close to the central binomial coefficient which led us to consider primes of the size $n/2 + O(\sqrt{n})$.

In this paper, we prove that an unconditional asymptotic formula for S_n holds for almost all n. We also establish an upper estimate of the sum. Then we show that a nontrivial lower estimate implies a bound for prime gaps $g_n \ll \sqrt{p_n}$. Previously, H. Cramér [C] proved that the Riemann Hypothesis (RH) implies $g_n \ll \sqrt{p_n} \log p_n$.

2. Experimental Observations

Throughout this work, we were guided by experimental data. Our initial version of the main theorem had $\frac{2^n}{\log n}$ as the dominant term of S_n, followed by an error term. We then performed calculations to analyze how the actual value of the sum S_n compared with the dominant term of its asymptotic by computing the ratio $\frac{S_n \log n}{2^n}$. As expected, this ratio was close to 1 with several observations either above or below 1.

An unexpected observation was that $\frac{S_n \log n}{2^n}$ appeared to have a bell-shaped distribution. However, there was one point of disagreement between theory and experimental data. While the data suggested that $\frac{S_n \log n}{2^n}$ has approximately bell-shaped distribution with a mean of about 1.06, our theory said that the mean value of $\frac{S_n \log n}{2^n}$ must approach 1 as $n \to \infty$. This disagreement between theory and experimental data led us to carefully reexamine both the theory and the data, and we found that we had oversimplified the dominant term of S_n which must actually be $\frac{2^n}{\log(n/2)}$ instead of $\frac{2^n}{\log n}$. With this small modification in the main theorem, both the theoretical and the experimental mean of $\frac{2^n}{\log(n/2)}$ converged towards the same value of 1. Also, this modification allowed us to improve the error term in the main theorem. The results of our computations are given below.

2.1. **Distribution of $\frac{S_n \log(n/2)}{2^n}$.** We computed the values of $\frac{S_n \log(n/2)}{2^n}$ to study its distribution. Using our available computing hardware, we were able to generate data for $n \leq 8.5 \times 10^4$ only. This is because as n increased in magnitude, the average run time to compute each incremental value of n was climbing and at around $n = 8.5 \times 10^4$, each n was taking a computing run time of 114 to 135 seconds. At this rate, even if we assume that there would be no further deceleration in computing speed, it would have taken us more than four years of nonstop computing to generate data for $n \leq 10^6$. Since the data generated thus far agreed with our theoretical derivations, it gave us the confidence that we were heading in the right direction and therefore we stopped computing at $n \leq 8.5 \times 10^4$.

The distribution of $\frac{S_n \log(n/2)}{2^n}$ (highlighted in red in the graph below) loosely resembled a bell curve. We used curve fitting to fit several curves to this distribution and observed that the best fit was obtained by a normal distribution (highlighted in blue in the graph below) with a mean of $\mu = 1$ with a 95% confidence interval range of (0.997128,1.003176) and standard deviation of $\sigma = 0.0932$ with a 95% confidence interval range of (0.090175,0.096224). For this fit, the coefficient of determination was $R^2 = 0.9642$ and the Akaike Information Criterion (AICc) value was 1473.84. We do not have a theoretical proof or disproof of normality. Based on the experimental observations, it is possible that the true distribution may approach normal as $n \to \infty$.

2.2. **Distribution over primes modulo a residue class.** Let $S_{n,a,b}$ be the sum of the binomial coefficients over all primes $p \leq n$ such that $p = ak + b$ for some positive integers a, b with $\gcd(a, b) = 1$, and k. Dirichlet's theorem for primes in arithmetic progression guarantees that as n increases, the number of primes in different residue classes for a given modulus are roughly equal. Hence, heuristically we expect that S_n is distributed roughly equally across all residue classes modulo a, i.e., $S_{n,a,b} \sim \frac{S_n}{\phi(a)} \sim \frac{2^n}{\phi(a)\log(n/2)}$. Further, if $\frac{S_n \log(n/2)}{2^n}$ has a certain distribution with a mean of 1, then we expect $\frac{S_{n,a,b} \log(n/2)}{2^n}$ to have a similar distribution with a mean of $\frac{1}{\phi(a)}$. We tested this hypothesis by computing the values of $S_{n,a,b}$ for different values of a and b and our experimental data supported the hypothesis. As an example, given below are the plots for the distribution of the binomial sum over primes of the form $12k + 1$, $12k + 5$, $12k + 7$, and $12k + 11$ shown in red, blue, green, and the black lines, respectively. As expected, the mean for each of these plots is about $\frac{1}{\phi(12)} = 0.25$.

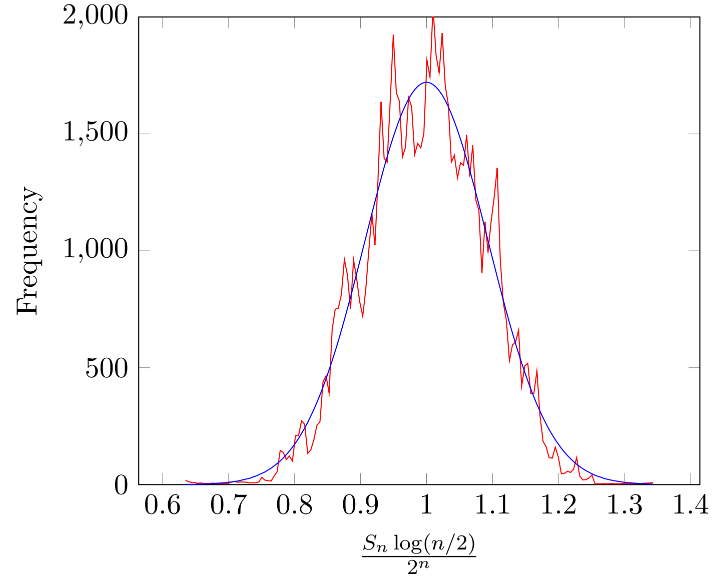

FIGURE 1. Distribution of $\frac{S_n \log(n/2)}{2^n}$

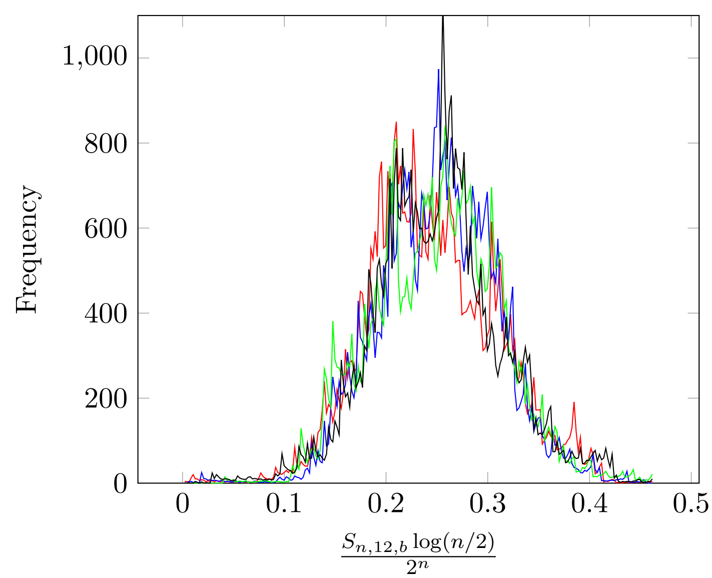

FIGURE 2. Distribution of $\frac{S_{n,12,b} \log(n/2)}{2^n}$

3. Experimental Data

The computed values of $\frac{S_n \log(n/2)}{2^n}$ for n ranging between 10×10^5 and 3.9×10^5 are given at every interval of 10^4 in the table below. All computations were programmed in Sagemath 8.1 and run on Intel i-7 8550U CPU 1.80GHz hardware.

n	$S_n \log(n/2)/2^n$	n	$S_n \log(n/2)/2^n$
100000	1.069169869	250000	0.986114371
110000	0.94301485	260000	0.965609639
120000	0.917190017	270000	0.973894862
130000	1.009817376	280000	0.99483856
140000	1.027465936	290000	0.953542586

150000	0.974742038	300000	1.028188428
160000	1.029105385	310000	0.993445284
170000	0.965422147	320000	1.017001058
180000	1.119848774	330000	0.869868372
190000	1.054380578	340000	1.073959735
200000	0.948608301	350000	0.873428088
210000	0.972819167	360000	1.090734815
220000	0.904355813	370000	1.024869577
230000	0.973834543	380000	0.965571714
240000	1.039784878	390000	1.025289725

4. Main Results

To explain the data, it is essential to enter deep into the error term of the asymptotic formula of S_n. The asymptotic formula is given for almost all n. The main ingredients are Huxley's zero density estimate [H], one of its consequences on primes in almost all short intervals [K, Theorem 7, Section 5.6], and Vinogradov's zero-free region for the Riemann zeta function. We find that the binomial coefficients $\binom{n}{p}$ for which $|p - n/2| \le \sqrt{N}\log N$ and $N \le n \le 2N$ mainly contribute to S_n.

Theorem 4.1. There is an absolute constant $c_0 > 0$ such that for almost all n,

$$S_n = \frac{2^n}{\log(n/2)} + O\left(2^n \exp\left(-c_0 \frac{(\log n)^{1/3}}{(\log\log n)^{1/3}}\right)\right) \text{ as } n \to \infty.$$

Here, *almost all* means that the number of $n \in [1, N] \cap \mathbb{Z}$ for which the asymptotic formula fails is $O(Ne^{-c_0(\log N)^{1/3}/(\log\log N)^{1/3}})$. The implied big-O constants are absolute. Note that the proportion of the exceptional set is approaching 0 as $N \to \infty$, but the exceptional set still contributes about 18% when we take $c_0 = 1$ and $N = 80000$. The exact value of c_0 is not evaluated here, but we have $c_0 < 1$ according to the proof in Section 5. We are currently not able to provide numerical observations for extreme large N such as 10^{200} due to the limitations of our computing hardware.

If we appeal to a zero density estimate [Mo, Theorem 12.1] that applies to Dirichlet L-functions, and the zero-free region for the Dirichlet L-functions [Mi], then we obtain the following generalization of Theorem 4.1.

Theorem 4.2. Let $(q, b) = 1$. There is an absolute constant $c_0 = c_0(q, b) > 0$ such that for almost all n,

$$S_{n,q,b} = \frac{2^n}{\phi(q)\log(n/2)} + O\left(2^n \exp\left(-c_0 \frac{(\log n)^{1/3}}{(\log\log n)^{1/3}}\right)\right) \text{ as } n \to \infty.$$

By Brun-Titchmarsh inequality, we are able to prove more than just the boundedness of $\frac{S_n \log(n/2)}{2^n}$. The constant in the upper bound $S_n \ll \frac{2^n}{\log n}$ can be refined and explicitly given.

Theorem 4.3. We have

$$\alpha := \liminf_{n\to\infty} \frac{S_n \log(n/2)}{2^n} \le 1 \le \limsup_{n\to\infty} \frac{S_n \log(n/2)}{2^n} \le 4.$$

The first two inequalities are by Theorem 4.1. The last one is achieved by a tighter use of Brun-Titchmarsh inequality. The numerical observation suggests that the upper bound would be 2 instead of 4. However, Brun-Titchmarsh inequality is not enough for proving this stronger upper bound.

On the other hand, we were unable to prove that $\alpha = \liminf \frac{S_n \log(n/2)}{2^n} > 0$ in this paper. We conjecture that $\alpha > 0$ and further that $S_n \sim \frac{2^n}{\log(n/2)}$ as $n \to \infty$. The values of $\frac{S_n \log(n/2)}{2^n}$ for some n up to $3.9 \cdot 10^5$ are provided in Section 3. Although proving $\alpha > 0$ was not successful, we found that it implies an unknown upper bound for prime gaps. We will show that the following theorem.

Theorem 4.4. The statement $\alpha > 0$ holds if and only if there are constants $b_1, b_2 > 0$ such that

$$\pi\left(\frac{n}{2} + b_1\sqrt{n}\right) - \pi\left(\frac{n}{2} - b_1\sqrt{n}\right) \geq \frac{b_2\sqrt{n}}{\log n} \text{ for all } n \geq N_0(b_1, b_2).$$

Thus, $\alpha > 0$ implies a bound for prime gaps $g_n \ll \sqrt{p_n}$ which is stronger than Cramér's bound $g_n \ll \sqrt{p_n}\log p_n$ conditional on the Riemann Hypothesis (see [C]). To see this, for sufficiently large n, consider $m \in \mathbb{N}$ with $p_{n-1} \leq m/2 - b_1\sqrt{m} < p_n$. Then we have $p_{n+1} \leq m/2 + b_1\sqrt{m}$. Thus, $p_{n+1} - p_n < 2b_1\sqrt{m} \leq C\sqrt{p_n}$.

Throughout this paper, we use the following notations.

- $\mathbf{P}(A)$ is the probability of an event A.
- $T_n \sim \mathrm{B}(n, \frac{1}{2})$ is the binomial distribution with n trials and the probability of success is $1/2$. Then we have $\mathbf{P}(T_n = k) = \binom{n}{k}/2^n$ for $0 \leq k \leq n$. T_n has the mean $n/2$, and the standard deviation $\sqrt{n}/2$.
- $\mathcal{P}$ is the set of prime numbers. Thus, $\mathbf{P}(T_n \in \mathcal{P}) = S_n/2^n$.
- $\pi(y) = \sum_{p\leq y} 1$ is the number of primes not exceeding y.
- $\psi(y) = \sum_{n\leq y} \Lambda(n)$ where Λ is the Von-Mangoldt function.
- $A(n) \ll B(n)$ means $|A(n)| \leq cB(n)$ for some positive absolute constant c.
- $\binom{x}{v} = \frac{\Gamma(x+1)}{\Gamma(v+1)\Gamma(x-v+1)} = \binom{x}{x-v}$ is the extension of binomial coefficients for real $x > 0$ and $v \geq 0$. For any $0 \leq v_1 \leq v_2 \leq x/2 \leq v_3 \leq v_4 \leq x$, we have $1 \leq \binom{x}{v_1} \leq \binom{x}{v_2} \leq \binom{x}{x/2} \geq \binom{x}{v_3} \geq \binom{x}{v_4} \geq 1$.
- $S_x = \sum_{p\leq x} \binom{x}{p}$ is an extension of S_n to positive real numbers.
- $S_{x,q,a} = \sum_{p\leq x,\ p\equiv a \bmod q} \binom{x}{p}$ is an extension of $S_{n,q,a}$ to positive real numbers.
- The letters j, k, n, p are integers. In particular, p denotes a prime. The letters $\alpha, \beta, \epsilon, t, v, x, X$ are real numbers. We write $c_0, c_1, c_2, \ldots$ for absolute positive constants.

5. Lemmas

We prove that the contribution of too large or too small primes to the sum S_x is negligible.

Lemma 5.1 (Hoeffding's Inequality). Let $X_1, \ldots, X_n$ be independent bounded random variables with $a \leq X_i \leq b$ for all i, and $\overline{X} = \frac{1}{n}\sum X_i$. Then for all $t \geq 0$,

$$\mathbf{P}\left(\left|\overline{X} - \mathbf{E}(\overline{X})\right| \geq t\right) \leq 2\exp\left(-\frac{2nt^2}{(b-a)^2}\right).$$

Applying this to an independent Bernoulli distribution with probability of success $1/2$, $T_n = \sum_{i\leq n} X_i = n\overline{X}$, and $t\sqrt{n} = h$, we have

$$\mathbf{P}\left(\left|T_n - \frac{n}{2}\right| \geq h\sqrt{n}\right) \leq 2e^{-2h^2}. \tag{1}$$

Corollary 5.1. For sufficiently large real $x > 0$ and $B_x = \left\{k \leq x : \left|k - \frac{x}{2}\right| \geq h\sqrt{x}\right\}$, we have

$$\frac{1}{2^x}\sum_{k\in\mathcal{P}\cap B_x}\binom{x}{k} \leq \frac{1}{2^x}\sum_{k\in B_x}\binom{x}{k} \leq 4e^{-2h^2}. \tag{2}$$

By Stirling's formula and $\log(1+t) = t - \frac{t^2}{2} + O(t^3)$ for $|t| \leq 1/2$, we have

Lemma 5.2. Let $g(x)$ be a function satisfying $|g(x)| \leq 6\log x$ and $x \to \infty$,

$$\frac{1}{2^x}\binom{x}{\frac{x}{2} + g(x)\sqrt{x}} = \frac{2}{\sqrt{2\pi x}}e^{-2(g(x))^2}\left(1 + O\left(\frac{(\log x)^3}{\sqrt{x}}\right)\right). \tag{3}$$

Proof. We apply Stirling's formula of the form:

$$\Gamma(x+1) = \sqrt{2\pi x}\left(\frac{x}{e}\right)^x\left(1 + O\left(\frac{1}{x}\right)\right).$$

Then we have

$$\begin{aligned}
&\binom{x}{\frac{x}{2}+g(x)\sqrt{x}}\\
&=\frac{\sqrt{2\pi x}\left(\frac{x}{e}\right)^x\left(1+O\left(\frac{1}{x}\right)\right)}{\sqrt{2\pi(\frac{x}{2}+g(x)\sqrt{x})}\left(\frac{\frac{x}{2}+g(x)\sqrt{x}}{e}\right)^{\frac{x}{2}+g(x)\sqrt{x}}\sqrt{2\pi(\frac{x}{2}-g(x)\sqrt{x})}\left(\frac{\frac{x}{2}-g(x)\sqrt{x}}{e}\right)^{\frac{x}{2}-g(x)\sqrt{x}}}\\
&=\frac{2}{\sqrt{2\pi x}}\frac{x^x\left(1+O\left(\frac{(\log x)^2}{x}\right)\right)}{\left(\frac{x}{2}+g(x)\sqrt{x}\right)^{\frac{x}{2}+g(x)\sqrt{x}}\left(\frac{x}{2}-g(x)\sqrt{x}\right)^{\frac{x}{2}-g(x)\sqrt{x}}}.
\end{aligned}$$

For the denominator, we apply $\log(1+t)=t-\frac{t^2}{2}+O(t^3)$ for $|t|\le 1/2$ repeatedly. The logarithm of the denominator satisfies

$$\begin{aligned}
&\log\left(\left(\frac{x}{2}+g(x)\sqrt{x}\right)^{\frac{x}{2}+g(x)\sqrt{x}}\left(\frac{x}{2}-g(x)\sqrt{x}\right)^{\frac{x}{2}-g(x)\sqrt{x}}\right)\\
&=\left(\frac{x}{2}+g(x)\sqrt{x}\right)\log\left(\frac{x}{2}+g(x)\sqrt{x}\right)+\left(\frac{x}{2}-g(x)\sqrt{x}\right)\log\left(\frac{x}{2}-g(x)\sqrt{x}\right)\\
&=x\log\left(\frac{x}{2}\right)+\left(\frac{x}{2}+g(x)\sqrt{x}\right)\log\left(1+\frac{2g(x)}{\sqrt{x}}\right)+\left(\frac{x}{2}-g(x)\sqrt{x}\right)\log\left(1-\frac{2g(x)}{\sqrt{x}}\right)\\
&=x\log\left(\frac{x}{2}\right)+\left(\frac{x}{2}+g(x)\sqrt{x}\right)\left(\frac{2g(x)}{\sqrt{x}}-\frac{1}{2}\left(\frac{2g(x)}{\sqrt{x}}\right)^2+O\left(\left(\frac{2g(x)}{\sqrt{x}}\right)^3\right)\right)\\
&\quad+\left(\frac{x}{2}-g(x)\sqrt{x}\right)\left(-\frac{2g(x)}{\sqrt{x}}-\frac{1}{2}\left(\frac{2g(x)}{\sqrt{x}}\right)^2+O\left(\left(\frac{2g(x)}{\sqrt{x}}\right)^3\right)\right)\\
&=x\log\left(\frac{x}{2}\right)+4(g(x))^2-2(g(x))^2+O\left(\frac{(g(x))^3}{\sqrt{x}}\right).
\end{aligned}$$

The result now follows. □

The following is the zero density estimate by Huxley [H].

Lemma 5.3 (Huxley 1972)**.** Given $0\le\sigma\le 1$ and $T\ge 2$, define

$$N(\sigma,T)=\left|\{\rho=\beta+i\gamma:\zeta(\rho)=0,\ \sigma\le\beta\le 1,|\gamma|\le T\}\right|.$$

There is an absolute constant $B>0$ such that

$$N(\sigma,T)\ll T^{2.4(1-\sigma)}(\log T)^B.$$

A version of results [K, Theorem 7, Section 5.6] on primes in almost all short intervals follows from the above. Denote by

$$L:=L(X)=\frac{(\log X)^{1/3}}{(\log\log X)^{1/3}}.$$

Corollary 5.2. Let $X^{-5/6+\epsilon}\le\delta\le X^{-1/6}$. There is an absolute positive constants $c_0:=c_0(\epsilon)>0$ and $X_0=X_0(\epsilon)>0$ such that for $x\in[X,2X]$, $X\ge X_0(\epsilon)$

$$\pi(x+\delta x)-\pi(x)=\frac{\delta x}{\log x}+O\left(\delta x e^{-c_0L}\right) \tag{4}$$

holds with an exceptional set of size at most $O(Xe^{-2c_0L})$.

Proof. The proof is along the same lines as [K, Theorem 7, Section 5.6], but the error terms are made stronger in this proof. Let

$$\theta(T):=\frac{b}{(\log T)^{2/3}(\log\log T)^{1/3}}. \tag{5}$$

The constant $b > 0$ in $\theta(T)$ is given by Vinogradov's zero-free region for the Riemann zeta function so that

$$\beta < 1 - \theta(T)$$

for any zeta-zero counted in $N(\sigma, T)$. Note that we can take $b = 1/57.54$ by [F]. Denote by $\mathcal{E}(X, \delta)$ the set of all $x \in [X, 2X]$ such that

$$|\psi(x + \delta x) - \psi(x) - \delta x| \geq \delta x e^{-c_1 L/2}.$$

We have

$$|\mathcal{E}(X, \delta)| \leq \int_X^{2X} (\delta x)^{-2} e^{c_1 L} |\psi(x + \delta x) - \psi(x) - \delta x|^2 \, dx.$$

Let $T = X^{5/6 - \epsilon/2}$. Then

$$\begin{aligned}
\psi(x + \delta x) - \psi(x) - \delta x &= \sum_{|\Im(\rho)| \leq T} \frac{(x + \delta x)^\rho - x^\rho}{\rho} + O(X^{1/6 + \epsilon/2} (\log X)^2) \\
&= \sum_{|\Im(\rho)| \leq T} x^\rho w(\rho) + O(X^{1/6 + 2\epsilon/3}),
\end{aligned}$$

where $w(\rho) = \int_1^{1+\delta} u^{\rho - 1} du$. By $|A + B|^2 \leq 2(|A|^2 + |B|^2)$ and $|w(\rho)| \leq \delta$, we have

$$\begin{aligned}
&\int_X^{2X} (\delta x)^{-2} e^{c_1 L} |\psi(x + \delta x) - \psi(x) - \delta x|^2 \, dx \\
&\ll X^{-2} e^{c_1 L} \int_X^{2X} \delta^{-2} \left(\left| \sum_{|\Im(\rho)| \leq T} x^\rho w(\rho) \right|^2 + O(\delta^2 x^{2 - \epsilon/2}) \right) dx \\
&\ll X^{-2} e^{c_1 L} \sum_{|\Im(\rho_1)| \leq T} \sum_{|\Im(\rho_2)| \leq T} \delta^{-2} w(\rho_1) \overline{w(\rho_2)} \int_X^{2X} x^{\rho_1 + \overline{\rho_2}} \, dx + e^{c_1 L} X^{1 - \epsilon/2} \\
&\ll X^{-2} e^{c_1 L} \sum_{|\Im(\rho_1)| \leq T} \sum_{|\Im(\rho_2)| \leq T} \left| \int_X^{2X} x^{\rho_1 + \overline{\rho_2}} \, dx \right| + e^{c_1 L} X^{1 - \epsilon/2}
\end{aligned}$$

Applying the inequality

$$\int_X^{2X} x^{\beta_1 + \beta_2 + i(\gamma_1 - \gamma_2)} \, dx \ll \frac{X^{\beta_1 + \beta_2 + 1}}{|\gamma_1 - \gamma_2| + 1},$$

the double sum is treated by Huxley's estimate (Lemma 5.3) as

$$\begin{aligned}
\sum_{|\Im(\rho_1)| \leq T} \sum_{|\Im(\rho_2)| \leq T} \left| \int_X^{2X} x^{\rho_1 + \overline{\rho_2}} \, dx \right| &\ll \sum_{|\gamma_1| \leq T} \sum_{|\gamma_2| \leq T} \frac{X^{\beta_1 + \beta_2 + 1}}{|\gamma_1 - \gamma_2| + 1} \\
&\ll \sum_{|\gamma_1| \leq T} \sum_{|\gamma_2| \leq T} \frac{X^{2\beta_1 + 1}}{|\gamma_1 - \gamma_2| + 1} \\
&\ll \sum_{|\gamma| \leq T} X^{2\beta + 1} (\log T)^2 \ll X (\log X)^2 \sum_{|\gamma| \leq T} X^{2\beta},
\end{aligned}$$

where the first inequality in the last line is due to

$$\sum_{|\gamma_2| \leq T} \frac{1}{|\gamma_1 - \gamma_2| + 1} \ll (\log T)^2$$

for any choice of γ_1. The sum is treated by partial summation.

$$\begin{aligned}\sum_{|\gamma|\le T} X^{2\beta} &= -\int_0^{1-\theta(T)} X^{2\sigma} dN(\sigma,T)\\ &\ll N(0,T) + \int_0^{1-\theta(T)} X^{2\sigma} N(\sigma,T) d\sigma\\ &\ll T(\log X) + (\log X)^B \int_0^{1-\theta(T)} X^{2\sigma} T^{2.4(1-\sigma)} d\sigma\\ &\ll X^2 e^{-c_2 L}.\end{aligned}$$

Here, $c_2 > 0$ is a constant which may depend on ϵ. We take $c_1 = c_2/2$. Then the result for $\psi(x)$ follows with $c_0 = c_1/2$.

Now we apply partial summation to obtain the result for $\pi(x)$. We write $\psi(x) = x + E(x)$ so that $E(x) = O(xe^{-c_4(\log x)^{3/5}/(\log\log x)^{1/5}})$. We have

$$\begin{aligned}&\pi(x+\delta x) - \pi(x)\\ &= \int_x^{x+\delta x} \frac{1}{\log t} d\psi(t) + O((\delta x)^{1/2})\\ &= \int_x^{x+\delta x} \frac{1}{\log t} dt + \frac{E(x+\delta x)}{\log(x+\delta x)} - \frac{E(x)}{\log x} + \int_x^{x+\delta x} \frac{E(t)}{t(\log t)^2}\, dt + O((\delta x)^{1/2})\\ &= \frac{\delta x}{\log x} + \frac{E(x+\delta x) - E(x)}{\log x} + O\left(\delta x \exp\left(-c_4 \frac{(\log X)^{3/5}}{(\log\log X)^{1/5}}\right)\right).\end{aligned}$$

For $x \in [X, 2X] - \mathcal{E}(X,\delta)$, we have

$$|E(x+\delta x) - E(x)| \le \delta x e^{-c_0 L}.$$

Then it follows that

$$\pi(x+\delta x) - \pi(x) = \frac{\delta x}{\log x} + O(\delta x e^{-c_0 L}).$$

□

The result is extended to multiple short intervals as follows.

Corollary 5.3. Let c_0 be the number in Corollary 5.2. Let $X \ge X_0$, $\delta = X^{-1/2} e^{-c_0 L}$, $h = \lfloor 5 e^{c_0 L} \log X \rfloor$, $x_0 := x_0(x) = \frac{x}{2} - \sqrt{X}\log X$, and $x_j := x_j(x) = (1+\delta)^j x_0$ for $j = 1, 2, \ldots, h$. Then there is a positive constant c_1 such that the set $\mathcal{E}(X)$ of all $x \in [X, 2X]$ for which

$$\left|\pi(x_{j+1}) - \pi(x_j) - \frac{x_{j+1} - x_j}{\log x_j}\right| \ge (x_{j+1} - x_j) e^{-c_0 L}$$

for some $j = 0, 1, 2, \ldots, h-1$ satisfies $\mu(\mathcal{E}(X)) \ll X e^{-c_1 L}$. Here, $\mu(A)$ is the Lebesgue measure of a set A.

Proof. For each j, we apply the method of Corollary 5.2 to prove that the set $\mathcal{E}_j(X)$ of $x \in [X, 2X]$ such that

$$\left|\pi(x_{j+1}) - \pi(x_j) - \frac{x_{j+1} - x_j}{\log x_j}\right| \ge (x_{j+1} - x_j) e^{-c_0 L}$$

satisfies $\mu(\mathcal{E}_j(X)) \ll X e^{-2c_0 L}$ uniformly for $j = 0, 1, 2, \ldots, h-1$. We take $\mathcal{E}(X) = \cup_{j=1}^h \mathcal{E}_j(X)$. Then the result follows by

$$\mu(\mathcal{E}(X)) \le \sum_{j=1}^h \mu(\mathcal{E}_j(X)) \ll X e^{-2c_0 L} e^{c_0 L} \log X.$$

□

Corollary 5.4. Under the same assumptions as in Corollary 5.3, for $X \geq X_0$, the set $\mathcal{E}(X)$ of all $n \in [X, 2X] \cap \mathbb{Z}$ for which

$$\left| \pi(x_{j+1}) - \pi(x_j) - \frac{x_{j+1} - x_j}{\log x_j} \right| \geq (x_{j+1} - x_j) e^{-c_0 L}$$

for some $j = 0, 1, 2, \ldots, h-1$ satisfies $|\mathcal{E}(X)| \ll X e^{-c_1 L}$. Here, $|A|$ is the cardinality of a set A.

For the similar results on primes in arithmetic progressions, we need a zero density estimate for Dirichlet L-functions. We need the following (see [Mo, Theorem 12.1]).

Lemma 5.4. Suppose that $q \geq 1$ and $T \geq 2$. Let $N(\sigma, T, \chi) = |\{\rho = \beta + i\gamma \ : \ L(s, \chi) = 0, \ \sigma \leq \beta \leq 1, \ |\gamma| \leq T\}$. For $\frac{1}{2} \leq \sigma \leq \frac{4}{5}$, we have

$$\sum_{\chi} N(\sigma, T, \chi) \ll (qT)^{\frac{3(1-\sigma)}{2-\sigma}} (\log qT)^9,$$

and for $\frac{4}{5} \leq \sigma \leq 1$, we have

$$\sum_{\chi} N(\sigma, T, \chi) \ll (qT)^{\frac{2(1-\sigma)}{\sigma}} (\log qT)^{14}.$$

Here, the sums are over all Dirichlet characters modulo q.

As a result, we have

$$\sum_{\chi} N(\sigma, T, \chi) \ll (qT)^{2.5(1-\sigma)} (\log qT)^{14}.$$

The following (see [Mi, Lemma 11]) is the zero-free region for the Dedekind zeta function. The zero-free regions for the Dirichlet L-functions follow from this.

Lemma 5.5 (Mitsui 1968)**.** Let $\zeta_K(s)$ be the Dedekind zeta function for a number field K. Then there is a positive constant c_K depending on K such that $\zeta_K(s)$ has no zeros in the region

$$\sigma \geq 1 - \frac{c_K}{(\log |t|)^{2/3} (\log \log |t|)^{1/3}}, \quad |t| \geq c_K.$$

Applying this lemma with $K = \mathbb{Q}(\zeta_q)$ and

$$\zeta_K(s) = \prod_{\chi \bmod q} L(s, \chi),$$

we see that there are no zeros of $L(s, \chi)$ in the above region for any χ modulo q.

Applying Lemma 5.4 and Lemma 5.5, we obtain the following analogue of Corollary 5.2.

Corollary 5.5. Let $X^{-4/5+\epsilon} \leq \delta \leq X^{-1/6}$, $(q, a) = 1$ and $A > 0$. There is an absolute positive constant $c_0 := c_0(\epsilon, A) > 0$ and $X_0 := X_0(\epsilon, A) > 0$ such that for $x \in [X, 2X]$, $X \geq X_0(\epsilon, A)$, and $q \leq (\log X)^A$,

$$\pi(x + \delta x; q, a) - \pi(x; q, a) = \frac{\delta x}{\phi(q) \log x} + O\left(\delta x e^{-c_0 L}\right) \tag{6}$$

holds with an exceptional set of size at most $O(X e^{-2c_0 L})$.

For the proof, we need to consider the possibility of the existence of the Landau-Siegel zero $\beta_1 \in \mathbb{R}$ of $L(s, \chi)$ with a real character χ modulo q. It is well-known that $\beta_1 < 1 - cq^{-\epsilon}$ for any $\epsilon > 0$ and a positive constant $c = c(q, \epsilon)$. In the proof of Corollary 5.2 where we treat the sum $\sum_{|\gamma| \leq T} X^{2\beta}$, the term $X^{2\beta_1}$ appears. This term is treated by $\beta_1 < 1 - cq^{-\epsilon}$ with a suitably chosen $\epsilon > 0$ so that $X^{2\beta_1} = O(X^2 \exp(-cL))$. If $q \leq (\log X)^A$ for some $A > 0$, we may choose $\epsilon = 1/(2A)$. Similarly, the following is an analogue of Corollary 5.3.

Corollary 5.6. Under the same assumptions as in Corollary 5.5, for $X \geq X_0$ and $q \leq (\log X)^A$, the set $\mathcal{E}(X)$ of all $n \in [X, 2X] \cap \mathbb{Z}$ for which

$$\left| \pi(x_{j+1}; q, a) - \pi(x_j; q, a) - \frac{x_{j+1} - x_j}{\phi(q) \log x_j} \right| \geq (x_{j+1} - x_j) e^{-c_0 L}$$

for some $j = 0, 1, 2, \ldots, h-1$ satisfies $|\mathcal{E}(X)| \ll X e^{-c_1 L}$. Here, $|A|$ is the cardinality of a set A.

6. Proof of Theorem 4.1 and Theorem 4.2

Let x_j and h be as in Corollary 5.3. Let $x \in [X, 2X] - \mathcal{E}(X)$ where $\mathcal{E}(X)$ is the set in Corollary 5.3 so that we can use

$$\left| \pi(x_{j+1}) - \pi(x_j) - \frac{x_{j+1} - x_j}{\log x_j} \right| \leq (x_{j+1} - x_j) e^{-c_0 L}, \tag{7}$$

for all $j = 0, 1, 2, \dots h-1$. By (2),

$$\frac{1}{2^x} \sum_{k \in \mathcal{P} \cap B_x} \binom{x}{k} \leq 4e^{-2(\log X)^2}, \tag{8}$$

where $B_x = \left\{ k \leq x : \left| k - \frac{x}{2} \right| \geq \sqrt{X} \log X \right\}$.

We treat S_x first over the intervals I_j and J defined as

$$I_j : \ (x_j, x_{j+1}] = \left(\frac{x}{2} + g(x_j)\sqrt{x}, \frac{x}{2} + g(x_{j+1})\sqrt{x} \right] \text{ for } j = 0, 1, 2, \dots h-1,$$

$$J : \ [0, x] - \bigcup_{j \leq h} I_j,$$

and $|g(x_j)| \leq 6 \log X$. Then we have by (3) and the mean value theorem,

$$\begin{aligned}
&\frac{1}{2^x} \sum_{p \in I_j} \binom{x}{p} \\
&= (\pi(x_{j+1}) - \pi(x_j)) \frac{2}{\sqrt{2\pi x}} e^{-2(g(x_j))^2} \left(1 + O\left(|g(x_j)| e^{-c_0 L}\right)\right) \left(1 + O\left(\frac{(\log X)^3}{\sqrt{x}}\right)\right) \\
&= \frac{x_{j+1} - x_j}{\log x_j} \frac{2}{\sqrt{2\pi x}} e^{-2(g(x_j))^2} \left(1 + O\left((\log X) e^{-c_0 L}\right)\right) \\
&= \frac{g(x_{j+1}) - g(x_j)}{\log x_j} \frac{2}{\sqrt{2\pi}} e^{-2(g(x_j))^2} \left(1 + O\left((\log X) e^{-c_0 L}\right)\right) \\
&= \frac{g(x_{j+1}) - g(x_j)}{\log(x/2)} \frac{2}{\sqrt{2\pi}} e^{-2(g(x_j))^2} \left(1 + O\left((\log X) e^{-c_0 L}\right)\right)
\end{aligned}$$

where the last equality is due to $\log x_j = \log(x/2)(1 + O(X^{-1/2}))$. The interval J contributes to

$$\frac{1}{2^x} \sum_{p \in J} \binom{x}{p} \leq 4e^{-2(\log X)^2}. \tag{9}$$

We now take the sum over $j \leq h$. Then we have

$$\left| \sum_{j \leq h} (g(x_{j+1}) - g(x_j)) \frac{2}{\sqrt{2\pi}} e^{-2(g(x_j))^2} - \int_{-\infty}^{\infty} \frac{2}{\sqrt{2\pi}} e^{-2u^2} \, du \right| \ll e^{-c_0 L}. \tag{10}$$

Putting together the sums over I_j's (8), and (9), we obtain

$$\left| \frac{1}{2^x} \sum_{p \leq x} \binom{x}{p} - \frac{1}{\log(x/2)} \right| \ll (\log X) e^{-c_0 L}.$$

Theorem 4.1 now follows by applying Corollary 5.4 instead of Corollary 5.3. Theorem 4.2 follows by applying Corollary 5.6.

7. Proof of Theorem 4.3

For the proof of Theorem 4.3 and 4.4, the results do not change if $\log(n/2)$ is replaced by $\log n$. The latter is more convenient for the proofs of Theorem 4.3 and 4.4. We prove the last inequality. Let $x \in [X, 2X]$, $h = \log X$, and $c = 1/\log X$. Let I_j's and J be defined by

$$I_j: \ \left(\frac{x}{2} + cj\sqrt{x}, \frac{x}{2} + c(j+1)\sqrt{x}\right] \text{ for } 0 \le |j| \le \frac{h}{c} \text{ and}$$

$$J: \ [0, x] - \bigcup_{0 \le |j| \le \frac{h}{c}} I_j.$$

We apply Brun-Titchmarsh theorem [MV, Corollary 3.4] on each I_j to obtain

$$\pi\left(\frac{x}{2} + c(j+1)\sqrt{x}\right) - \pi\left(\frac{x}{2} + cj\sqrt{x}\right) \le \frac{2c\sqrt{x}}{\log(c\sqrt{x})}\left(1 + O\left(\frac{1}{\log x}\right)\right) \tag{11}$$

$$= \frac{4c\sqrt{x}}{\log x}\left(1 + O\left(\frac{\log\log x}{\log x}\right)\right).$$

Then the upper bound of the sum over I_j is given by

$$\frac{1}{2^x}\sum_{p \in I_j}\binom{x}{p} \le \frac{4c\sqrt{x}}{\log x}\left(1 + O\left(\frac{\log\log x}{\log x}\right)\right)\frac{2}{\sqrt{2\pi x}}e^{-2(cr_j)^2}, \tag{12}$$

where $e^{-2(cr_j)^2} = \max\limits_{x \in [cj, c(j+1)]} e^{-2x^2}$. Summing over $|j| \le h/c$ and applying (12), we obtain

$$\frac{S_x}{2^x} \le \frac{1}{\log x}\left(4 + O\left(\frac{\log\log x}{\log x}\right)\right) \tag{13}$$

It follows that

$$\limsup_{n \to \infty}\frac{S_n \log n}{2^n} \le \limsup_{x \to \infty}\frac{S_x \log x}{2^x} \le 4.$$

8. Proof of Theorem 4.4

Let us first assume $0 < \alpha = \liminf_{n \to \infty}\frac{S_n \log n}{2^n}$. Then for sufficiently large n,

$$\frac{11\alpha}{12} \le \frac{S_n \log n}{2^n}.$$

Let $c = 1/\log n$ and $h = \log n$. Then by Hoeffding's inequality,

$$\mathbf{P}\left(T_n \in \mathcal{P}, \left|T_n - \frac{n}{2}\right| \ge h\sqrt{n}\right) \le \mathbf{P}\left(\left|T_n - \frac{n}{2}\right| \ge h\sqrt{n}\right) \le 2e^{-2h^2} = 2e^{-2(\log n)^2}.$$

We use the subintervals I_j and J for $|j| \le c/h$ as follows. These subintervals are defined by

$$I_j: \ \left(\frac{x}{2} + cj\sqrt{x}, \frac{x}{2} + c(j+1)\sqrt{x}\right] \text{ for } 0 \le |j| \le \frac{h}{c} \text{ and}$$

$$J: \ [0, x] - \bigcup_{0 \le |j| \le \frac{h}{c}} I_j.$$

Then we have

$$\sum_{p \in J}\binom{n}{p} \ll 2^n e^{-2(\log n)^2}.$$

Apply the Brun-Titchmarsh inequality and choose $b_1 > 0$ so that the contribution of primes in the intervals I_j with $b_1 \le c|j| \le h$ is bounded by

$$\sum_{b_1 \le c|j| \le h}\sum_{p \in I_j}\binom{n}{p} \le \frac{2^n}{\log n}\left(\int_{|t| \ge b_1}\frac{2}{\sqrt{2\pi}}e^{-2t^2}\,dt + O\left(\frac{1}{\log n}\right)\right) \le \frac{2^n\alpha}{2\log n}.$$

For example, let b_1 satisfy $\int_{|t|\geq b_1} e^{-t^2} dt < \alpha\sqrt{2\pi}/6$. Then the contribution of primes in the interval $n/2 - b_1\sqrt{n} < p \leq n/2 + b_1\sqrt{n}$ is bounded below by $\frac{2^n \alpha}{3\log n}$ for sufficiently large n. Thus,

$$\binom{n}{n/2}\left(\pi\left(\frac{n}{2}+b_1\sqrt{n}\right)-\pi\left(\frac{n}{2}-b_1\sqrt{n}\right)\right) \geq \frac{2^n\alpha}{3\log n}.$$

By Lemma 5.2, we have

$$\frac{2}{\sqrt{2\pi n}}\left(\pi\left(\frac{n}{2}+b_1\sqrt{n}\right)-\pi\left(\frac{n}{2}-b_1\sqrt{n}\right)\right) \geq \frac{\alpha}{3\log n}\left(1+O\left(\frac{h^3}{\sqrt{n}}\right)\right).$$

Now this yields the lower bound for the number of primes in the short interval. There is $b_2 = \alpha\sqrt{2\pi}/6 > 0$ such that for $n \geq N_0$,

$$\pi\left(\frac{n}{2}+b_1\sqrt{n}\right)-\pi\left(\frac{n}{2}-b_1\sqrt{n}\right) \geq \frac{b_2\sqrt{n}}{\log n}.$$

For the converse, assume that there are $b_1, b_2 > 0$ such that for $n \geq N_0$, we have

$$\pi\left(\frac{n}{2}+b_1\sqrt{n}\right)-\pi\left(\frac{n}{2}-b_1\sqrt{n}\right) \geq \frac{b_2\sqrt{n}}{\log n}.$$

Then by Lemma 5.2, there is an absolute constant $b_3 > 0$ such that,

$$\begin{aligned} S_n \geq \sum_{\left|\frac{n}{2}-p\right|\leq b_1\sqrt{n}} \binom{n}{p} &\geq \binom{n}{\frac{n}{2}+b_1\sqrt{n}}\left(\pi\left(\frac{n}{2}+b_1\sqrt{n}\right)-\pi\left(\frac{n}{2}-b_1\sqrt{n}\right)\right) \\ &\geq 2^n\frac{2}{\sqrt{2\pi n}}e^{-2b_1^2}\frac{b_2\sqrt{n}}{\log n}\left(1+O\left(\frac{h^3}{\sqrt{n}}\right)\right) \geq \frac{2^n b_3}{\log n}. \end{aligned}$$

Here, we may take $b_3 = \frac{2}{\sqrt{2\pi}}e^{-2b_1^2}b_2$. Therefore, Theorem 4.4 follows.

9. Remarks

This work began out of the curiosity of understanding how the asymptotic of the sum $\sum_{a_r\leq n}^{n}\binom{n}{a_r}$ would behave over the most interesting subset of natural numbers, the set of prime numbers $a_r = p_r$. However, during the course of our work, we expanded our subset of natural numbers to investigate the asymptotic growth rate of the binomial sum over the set of squares, the set of natural numbers co-prime to n, etc. We also investigated the product of the binomial coefficients over different subsets of natural numbers as described above. We believe that our methods can be applied to explore other properties of binomial sums. Of these, we found the summation over squares to be the most interesting and perhaps could be of some practical importance in physics as it displayed properties analogous to the Fourier series expansion of the heat equation. Hence, we have kept our work on the binomial sum over squares outside the scope of our current paper to give it a separate treatment in an independent paper of its own right.

References

[C] H. Cramér, On the distribution of primes, *Proc. Camb. Phil. Soc.* **20**, (1920), 272-280.

[F] K. Ford, Zero Free Regions for the Riemann Zeta Function, *Number Theory for the Millenium (Urbana, IL, 2000) (M. A. Bennett, B. C. Berndt, N. Boston, H. G. Diamond, A. J. Hildebrand and W. Phillip, eds)*, **2**, (2002), 25-56.

[H] M. N. Huxley, On the Difference between Consecutive Primes, *Inventiones Math*, **15**, (1972), 164-170.

[K] A. Kumchev, *The Distribution of Prime Numbers*, Lecture Notes, 2005. `https://tigerweb.towson.edu/akumchev/distributionofprimesnotes.pdf`

[Mi] T. Mitsui, On the Prime Ideal Theorem, *J. Math. Soc. Japan*, **20**, (1968), 233-247.

[Mo] H. Montgomery, *Topics in Multiplicative Number Theory*, Springer, 1971.

[MV] H. Montgomery, R. Vaughan, *Multiplicative Number Theory I. Classical Theory*, Cambridge University Press, 2007.

[S] N. K. Sinha, *What is the sum of the binomial coefficients* $\binom{n}{p}$ *over prime numbers?*, `https://math.stackexchange.com/q/2930593`, Mathematics Stack Exchange.

(Sungjin Kim) Santa Monica College, California State University Northridge
E-mail address: 707107@gmail.com

(Nilotpal Kanti Sinha) Department of Culture and Tourism, Abu Dhabi, UAE
E-mail address: nilotpalsinha@gmail.com