#A54 INTEGERS 17 (2017)

WHEN ALMOST ALL SETS ARE DIFFERENCE DOMINATED IN $\mathbb{Z}/n\mathbb{Z}$

Anand Hemmady[1]
Department of Mathematics and Statistics, Williams College, Williamstown, Massachusetts
ash6@williams.edu

Adam Lott
Department of Mathematics, University of Rochester, Rochester, New York
alott@u.rochester.edu

Steven J. Miller
Department of Mathematics and Statistics, Williams College, Williamstown, Massachusetts
sjm1@williams.edu, Steven.Miller.Mc.96@aya.yale.edu

Received: 8/11/16, Revised: 8/26/17, Accepted: 10/30/17, Published: 11/10/17

Abstract

We investigate the behavior of the sum and difference sets of $A \subseteq \mathbb{Z}/n\mathbb{Z}$ chosen independently and randomly according to a binomial parameter $p(n) = o(1)$. We show that for rapidly decaying $p(n)$, A is almost surely difference-dominated as $n \to \infty$, but for slowly decaying $p(n)$, A is almost surely balanced as $n \to \infty$, with a continuous phase transition as $p(n)$ crosses a critical threshold. Specifically, we show that if $p(n) = o(n^{-1/2})$, then $|A - A|/|A + A|$ converges to 2 almost surely as $n \to \infty$ and if $p(n) = c \cdot n^{-1/2}$, then $|A - A|/|A + A|$ converges to $1 + \exp(-c^2/2)$ almost surely as $n \to \infty$. In these cases, we modify the arguments of Hegarty and Miller on subsets of $\mathbb{Z}$ to prove our results. When $\sqrt{\log n} \cdot n^{-1/2} = o(p(n))$, we prove that $|A - A| = |A + A| = n$ almost surely as $n \to \infty$ if some additional restrictions are placed on n. In this case, the behavior is drastically different from that of subsets of $\mathbb{Z}$ and new technical issues arise, so a novel approach is needed. When $n^{-1/2} = o(p(n))$ and $p(n) = O(\sqrt{\log n} \cdot n^{-1/2})$, the behavior of $|A + A|$ and $|A - A|$ is markedly different and suggests an avenue for further study. These results establish a "correspondence principle" with the existing results of Hegarty, Miller, and Vissuet. As $p(n)$ decays more rapidly, the behavior of subsets of $\mathbb{Z}/n\mathbb{Z}$ approaches the behavior of subsets of $\mathbb{Z}$ shown by Hegarty and Miller. Moreover, as $p(n)$ decays more slowly, the behavior of subsets of $\mathbb{Z}/n\mathbb{Z}$ approaches the behavior shown by Miller and Vissuet in the case where $p(n) = 1/2$.

[1] This work was supported by NSF Grants DMS1265673, DMS1561945, and DMS1347804, Williams College, and the Williams College Finnerty Fund. We also thank Oleg Lazarev and Kevin O'Bryant for the TikZ code used to generate the images in this work.

1. Introduction

A central object of study in additive combinatorics is the sumset of a set. Given an abelian group G (written additively) and a set $A \subseteq G$, we define its sumset $A+A := \{a+b : a, b \in A\}$. Similarly, we can define its difference set $A - A := \{a-b : a, b \in A\}$. If $|A+A| > |A-A|$, we say A is *sum-dominated* or a *More Sums Than Differences (MSTD) set.* If $|A-A| > |A+A|$, we say A is *difference-dominated*, and if $|A+A| = |A-A|$ we say A is *balanced*. The most common setting for studying MSTD sets is subsets of $\mathbb{Z}$ (though they have been studied elsewhere as well; see, for example, [9] and [2]). Since addition in $\mathbb{Z}$ is commutative but subtraction is not, we typically expect most sets to be difference-dominated. As Nathanson [10] famously remarked,

> "Even though there exist sets A which have more sums than differences, such sets should be rare, and it must be true with the right way of counting that the vast majority of sets satisfies $|A-A| > |A+A|$."

Surprisingly, Martin and O'Bryant [8] showed that a positive proportion of subsets of $\{0, \ldots, n-1\} \subset \mathbb{Z}$ are sum-dominated in the limit as $n \to \infty$. Zhao [11] has shown that this proportion is around 4.5×10^{-4}.

Martin and O'Bryant proved their result by picking sets $A \subseteq \{0, \ldots, n-1\} \subset \mathbb{Z}$ randomly according to a binomial parameter $p = 1/2$ (i.e., every subset is equally likely) and showing that the probability of being sum-dominated is nonzero as $n \to \infty$. This happens because if A is large enough, almost all possible sums and differences appear, so it is possible to choose A carefully to be sum-dominated. However, Hegarty and Miller [3] showed that if $A \subseteq \{0, \ldots, n-1\} \subset \mathbb{Z}$ is instead picked randomly according to a binomial parameter $p(n) = o(1)$, then the probability of being sum-dominated tends to 0 as $n \to \infty$. In some sense, this is Nathanson's "right way of counting" because it prevents A from being too large.

In this paper, we examine subsets of $\mathbb{Z}/n\mathbb{Z}$. Miller and Vissuet [9] showed that if subsets of $\mathbb{Z}/n\mathbb{Z}$ are picked uniformly at random, then they are balanced with probability 1 as $n \to \infty$. In the style of [3], we instead pick subsets randomly according to a binomial parameter $p(n) = o(1)$. Our main result is the following.

Theorem 1.1. *Let $A \subseteq \mathbb{Z}/n\mathbb{Z}$ be a subset chosen randomly according to a binomial parameter $p(n) = o(1)$. Let S, D denote the random variables $|A+A|$, $|A-A|$ respectively. We have three cases.*

1. *Fast decay:*
   If $p(n) = o(n^{-1/2})$, then

   (a) $S \sim \frac{1}{2}(n \cdot p(n))^2$,

   (b) $D \sim (n \cdot p(n))^2$.

2. *Critical decay:*
   If $p(n) = cn^{-1/2}$, *then*

   (a) $S \sim n(1 - \exp(-c^2/2))$,

   (b) $D \sim n(1 - \exp(1 - c^2))$.

3. *Slow decay:*
   If $\sqrt{\log n} \cdot n^{-1/2} = o(p(n))$ *and* n *is prime, then*

   (a) $S \sim n$,

   (b) $D \sim n$.

Remark 1.2. Throughout, we will point out instances where the case $n^{-1/2} = o(p(n))$ and $p(n) = O(\sqrt{\log n} \cdot n^{-1/2})$ causes deviant behavior.

Remark 1.3. In part 3 we assume that n is prime to simplify the technical details of our analysis; however, numerical simulations suggest that the behavior is the same for any n.

For parts (1) and (2) of Theorem 1.1, we modify the arguments in [3] to work in this new environment where sums and differences are considered modulo n; however, for part 3, these methods do not work and a new approach is needed.

We first fix some notation.

- If X is a random variable depending on n, we write $X \sim f(n)$ if for every $\epsilon > 0$, $\text{Prob}((1-\epsilon)f(n) < X < (1+\epsilon)f(n)) \to 1$ as $n \to \infty$.
- If X and Y are two quantities depending on n, we also write $X \sim Y$ if $\lim_{n\to\infty} X/Y = 1$. This abuse of notation should not cause any confusion as it will be clear from context if we are talking about a random variable or not.
- We say $f(n) = O(g(n))$ if $\limsup_{n\to\infty} f(n)/g(n) < \infty$, and we say $f(n) = o(g(n))$ if $\lim_{n\to\infty} f(n)/g(n) = 0$.
- To reduce clutter, we write p in place of $p(n)$ and the dependence on n is implied.

2. Proof of Main Result: Fast and Critical Decay Cases

To prove parts 1 and 2 of Theorem 1.1, we show that the expected value of each random variable is as claimed, and then show that the variable is strongly concentrated about its mean.

We use the following construction from [3]. Let

$$X_k = \#\{\{\{a_1, a_2\}, \ldots, \{a_{2k-1}, a_{2k}\}\} : a_i \in A,\ a_1 + a_2 = \ldots = a_{2k-1} + a_{2k}\} \text{ and} \tag{2.1}$$

$$Y_k = \#\{\{(a_1, a_2), \ldots, (a_{2k-1}, a_{2k})\} : a_i \in A,\ a_1 - a_2 = \ldots = a_{2k-1} - a_{2k}\}. \tag{2.2}$$

In words, X_k denotes the number of times k pairs of elements from A all have the same sum, and Y_k denotes the number of times k pairs of elements from A all have the same difference. It is important to note that X_k consists of *unordered* pairs of elements, while Y_k consists of *ordered* pairs. Since A is a randomly chosen set, X_k and Y_k are random variables. The idea is that X_k and Y_k measure the number of repeated sums and differences, so if we can control these quantities, we can control $|A + A|$ and $|A - A|$. We have the following lemma.

Lemma 2.1. *If $p(n) = O(n^{-1/2})$, then*

1. $X_k \sim \frac{n^{k+1}}{k!}\left(\frac{p^2}{2}\right)^k$, *and*
2. $Y_k \sim \frac{n^{k+1}}{k!}(p^2)^k$.

Proof. Each k-tuple that contributes to X_k is one of two types: either all $2k$ elements are distinct, or one of the pairs is a repeated element. Following the notation of [3], let ξ_{1k}, ξ_{2k} be the number of tuples of the first type and second type, respectively. Since every element of A has $\lceil n/2 \rceil$ representations[2] as the sum of two elements of A, we have

$$\xi_{1k} = \sum_{r=0}^{n-1} \binom{\lceil n/2 \rceil}{k} = n\binom{\lceil n/2 \rceil}{k} \sim n\frac{(n/2)^k}{k!} \sim \frac{n^{k+1}}{2^k k!} \tag{2.3}$$

$$\xi_{2k} = \sum_{r=0}^{n-1} \binom{\lceil n/2 \rceil}{k-1} = n\binom{\lceil n/2 \rceil}{k-1} \sim n\frac{(n/2)^{k-1}}{(k-1)!} \sim \frac{n^k}{2^{k-1}(k-1)!}. \tag{2.4}$$

The expected value of X_k is then given by

$$\mathbb{E}[X_k] = \xi_{1k}p^{2k} + \xi_{2k}p^{2k-1} = \frac{n^{k+1}}{2^k k!}p^{2k} + \frac{n^k}{2^{k-1}(k-1)!}p^{2k-1} \sim \frac{n^{k+1}}{k!}\left(\frac{p^2}{2}\right)^k. \tag{2.5}$$

Now we show that the variance of X_k is small enough to guarantee strong concentration about the mean. It is sufficient to show that $\mathrm{Var}(X_k) = o(\mathbb{E}[X_k]^2)$ (see, for

[2] Note that this is the fundamental difference between considering sums in the normal sense and considering sums mod n. In the regular setting, the number of representations of k as a sum depends on k, but in this setting it does not. This difference is what causes the different constants in part (2) of Theorem 1.1.

example, chapter 4 of [1]). We have

$$\mathrm{Var}(X_k) = \sum_{\alpha} \mathrm{Var}(Y_\alpha) + \sum_{\alpha\neq\beta} \mathrm{Cov}(Y_\alpha, Y_\beta)\,, \tag{2.6}$$

where the sums are over k-tuples of unordered pairs of elements of A and Y_α is an indicator variable that equals 1 if α contributes to X_k and 0 otherwise. From the arguments in [1], it is enough to show that

$$\sum_{\alpha,\beta} \mathrm{Prob}\,(\alpha, \beta \text{ both contribute to } X_k) = o(\mathbb{E}[X_k]^2), \tag{2.7}$$

where the sum is now over all α, β that have at least one member in common. The main contribution to this sum comes from pairs α, β with one element in common and $2k$ distinct elements each, and there are $O(n^{2k+1})$ choices for this (see the proof of Lemma 2.1 in [3] for details). Thus the sum (2.7) is at most $O(n^{2k+1}p^{4k-1}) = o(n^{2k+2}p^{4k})$. Thus part 1 is proven.

The proof of part 2 follows the exact same argument, so we omit the details. □

We can now prove parts (1) and (2) of Theorem 1.1.

Proof of Theorem 1.1, part (1).
If $p(n) = o(n^{-1/2})$, we have by Lemma 2.1 that $X_1 \sim \frac{1}{2}(n \cdot p(n))^2$, $Y_1 \sim (n \cdot p(n))^2$, $X_k = o(X_1)$, and $Y_k = o(Y_1)$ for $k \geq 2$. In other words, all but a vanishing proportion of pairs of elements in A have distinct sums and differences. Thus $S \sim \frac{1}{2}(n \cdot p(n))^2$ and $D \sim (n \cdot p(n))^2$ as claimed. This proves part (1). □

Proof of Theorem 1.1, part (2).
By inclusion-exclusion, we have that

$$S \;=\; \sum_{k=1}^{\infty} (-1)^{k+1} X_k. \tag{2.8}$$

Lemma 2.1 yields $X_k \sim n\frac{1}{k!}\left(\frac{c^2}{2}\right)^k$, so (2.8) gives

$$S \;\sim\; n \cdot \sum_{k=1}^{\infty} \frac{(-1)^k}{k!}\left(\frac{c^2}{2}\right)^k \;=\; n(1 - \exp(-c^2/2)), \tag{2.9}$$

which was the claim. Similarly, for differences we have

$$\begin{aligned} D \;&=\; \sum_{k=1}^{\infty} (-1)^{k+1} Y_k \quad \text{and} \\ Y_k \;&\sim\; n\frac{1}{k!}(c^2)^k, \end{aligned} \tag{2.10}$$

so

$$D \ \sim \ n \cdot \sum_{k=1}^{\infty} \frac{(-1)^k}{k!}(c^2)^k \ = \ n(1-\exp(-c^2)). \tag{2.11}$$

This proves part (2). □

3. Proof of Main Result: Slow Decay Case

We need the following bound.

Lemma 3.1. *Suppose $p(n) = n^{-\delta}$ where $\delta \in (0, 1/2)$. Let*

$$F(n) \ = \ \sum_{r=0}^{n/2} \binom{n-r}{r} p^r (1-p)^{n-r}. \tag{3.1}$$

Then $F(n) = o(1/n^3)$.

This is proven in Appendix A.

To prove part 3 of Theorem 1.1, we use the following strategy. We let $S^c = n - |A+A|$ be the number of sums missing from $A+A$, and we show that

$$\lim_{n\to\infty} \mathbb{E}[S^c] \ = \ \lim_{n\to\infty} \mathrm{Var}(S^c) \ = \ 0. \tag{3.2}$$

To show that this is sufficient, let $v(n) = \mathrm{Var}(S^c)$ and let $s(n) = \sqrt{v(n)}$. By Chebyshev's inequality

$$\mathrm{Prob}\left(|S^c - \mathbb{E}[S^c]| \ \geq \ ks(n)\right) \ \leq \ \frac{1}{k^2}. \tag{3.3}$$

Taking $k = 1/\sqrt{s(n)}$, we see that

$$\mathrm{Prob}\left(|S^c - \mathbb{E}[S^c]| \ \geq \ \sqrt{s(n)}\right) \ \leq \ s(n). \tag{3.4}$$

Thus, since $\mathbb{E}[S^c]$ also tends to 0, we can say that $\mathrm{Prob}\left(S^c > 1/2\right) \to 0$ as $n \to \infty$; thus $S \sim n$. We also use this argument for differences by replacing S^c everywhere with $D^c := n - |A-A|$. We can now prove part (3) of Theorem 1.1.

Proof of Theorem 1.1, part (3a).
Let $S^c = n - |A+A|$. First we compute $\mathbb{E}[S^c]$. Define the random variables Z_k by

$$Z_k \ := \ \begin{cases} 1, & k \notin A+A \\ 0, & k \in A+A \end{cases} \tag{3.5}$$

so that $\sum_{k\in\mathbb{Z}/n\mathbb{Z}} Z_k = S^c$.

Since n is assumed to be a large prime and is therefore odd, each $k \in \mathbb{Z}/n\mathbb{Z}$ can be written as a sum in $(n+1)/2$ different ways, and all of the representations are independent of each other, so $\text{Prob}\,(k \notin A+A) = \mathbb{E}[Z_k] = (1-p^2)^{(n+1)/2}$. Thus we have

$$\mathbb{E}[S^c] \;=\; \sum_{k \in \mathbb{Z}/n\mathbb{Z}} \mathbb{E}[Z_k] \;=\; n(1-p^2)^{(n+1)/2} \;\sim\; n(1-p^2)^{n/2}. \tag{3.6}$$

Denote this quantity by $G(n)$. To show that it tends to 0, we have

$$\begin{aligned} \log G(n) &= \log n + \frac{1}{2} n \log(1-p^2) \\ &= \log n + \frac{1}{2} n(-p^2 + O(p^4)) \\ &= \log n - \frac{1}{2} np^2 + O(np^4), \end{aligned} \tag{3.7}$$

which tends to $-\infty$ as $n \to \infty$ because $\log n = o(np^2)$; thus $G(n)$ tends to 0.

Remark 3.2. If instead we had $p(n) = o(\sqrt{\log n} \cdot n^{-1/2})$, then $\log G(n)$ would tend to $+\infty$ rather than $-\infty$.

We now compute $\text{Var}(S^c)$. We have

$$\begin{aligned} \text{Var}(S^c) &= \sum_{k \in \mathbb{Z}/n\mathbb{Z}} \text{Var}(Z_k) + \sum_{i \neq j \in \mathbb{Z}/n\mathbb{Z}} \text{Cov}(Z_i, Z_j) \\ &= \sum_k \left(\mathbb{E}[Z_k^2] - \mathbb{E}[Z_k]^2\right) + \sum_{i \neq j} \left(\mathbb{E}[Z_i Z_j] - \mathbb{E}[Z_i]\mathbb{E}[Z_j]\right) \\ &\sim \sum_k \Big((1-p^2)^{n/2} - (1-p^2)^n\Big) \\ &\quad + \sum_{i \neq j} \left(\text{Prob}\,(i \notin A+A \;\wedge\; j \notin A+A) - (1-p^2)^n\right) \\ &\sim n(1-p^2)^{n/2} - n^2(1-p^2)^n + \sum_{i \neq j} \text{Prob}\,(i \notin A+A \;\wedge\; j \notin A+A)\,. \end{aligned} \tag{3.8}$$

We can get an expression for the probability that i and j are both missing from the sumset by translating the problem into graph theory. Define the graph $G^S_{n,i,j}$ as follows. The vertices of $G^S_{n,i,j}$ are the elements $\{0, \ldots, n-1\}$, and vertices a and b are connected by an edge if and only if $a + b \equiv i \pmod{n}$ or $a + b \equiv j \pmod{n}$ (see Figure 1).

The event $(i \notin A+A \wedge j \notin A+A)$ corresponds to the event that no two adjacent vertices of $G^S_{n,i,j}$ are in A. Since we have assumed n is prime, we know that for any i, j, $G^S_{n,i,j}$ is isomorphic to a path of n vertices with a loop on each endpoint (see Figure 2).

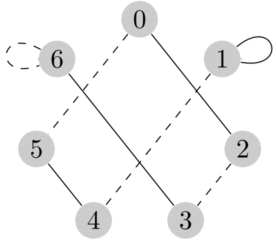

Figure 1: The graph $G^S_{7,2,5}$. For clarity, each edge is solid or dotted depending on the sum of the two incident vertices, but this doesn't affect the graph.

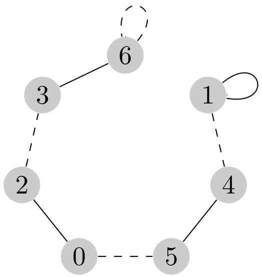

Figure 2: The graph $G^S_{7,2,5}$ from Figure 1 rearranged to illustrate the structure. The graph $G^S_{n,i,j}$ has this structure for any n, i, j.

We see that A cannot contain either of the two endpoints (6 and 1 in the figure). So, after a relabeling of the vertices, picking a set A so that i and j are both missing from $A + A$ is equivalent to picking a subset of $\{1, \ldots, n-2\}$ with no two consecutive elements (1 and $n-2$ are not considered consecutive). Since we are picking elements of A independently with probability p, the probability of picking A with no two consecutive elements is

$$\sum_{r=0}^{(n-2)/2} C(n-2, r) p^r (1-p)^{(n-2)-r}, \tag{3.9}$$

where $C(n-2, r)$ denotes the number of r-element subsets of $\{1, \ldots, n-2\}$ with no consecutive elements. By a simple counting argument (see the calculation of quantity Y in Appendix B), we have $C(n-2, r) = \binom{n-2-r+1}{r}$.

Remark 3.3. The numbers $C(n-2, r)$ also have another combinatorial interpretation. Any positive integer can be written uniquely as a sum of non-adjacent Fibonacci summands; these numbers are how many integers at most $F_{n-1} - 1$ have exactly r summands. This partition of the integers in $[0, F_{n-1} - 1)$ was used in [5] to show that the distribution of the number of summands converges to a Gaussian as $n \to \infty$.

Since the probability that neither of the endpoints gets picked is $(1-p)^2$, we have that

$$\begin{aligned}\text{Prob}\,(i \notin A+A \ \wedge \ j \notin A+A) \ &= \ (1-p)^2 \sum_{r=0}^{(n-2)/2} \binom{n-2-r+1}{r} p^r (1-p)^{(n-2)-r} \\ &\leq \ \sum_{r=0}^{n/2} \binom{n-r}{r} p^r (1-p)^{n-r}. \qquad (3.10)\end{aligned}$$

Recall that (3.10) is the quantity $F(n)$ from Lemma 3.1. So we have

$$\begin{aligned}\text{Var}(S^c) \ &\leq \ n(1-p^2)^{n/2} - n^2(1-p^2)^n + \sum_{i \neq j} F(n) \\ &\leq \ n(1-p^2)^{n/2} - n^2(1-p^2)^n + n^2 F(n). \qquad (3.11)\end{aligned}$$

The first term is $\mathbb{E}[S^c]$, which tends to 0. The second term is $\mathbb{E}[S^c]^2$, which also tends to 0. The third term tends to 0 by Lemma 3.1, so $\text{Var}(S^c)$ tends to 0 as $n \to \infty$. This completes the proof that $S \sim n$. □

Proof of Theorem 1.1, part (3b).
We let $D^c := n - |A - A|$, so D^c denotes the number of differences missing from $A - A$. We will compute $\mathbb{E}[D^c]$ and $\text{Var}(D^c)$ and show that

$$\lim_{n\to\infty} \mathbb{E}[D^c] \ = \ \lim_{n\to\infty} \text{Var}(D^c) \ = \ 0. \qquad (3.12)$$

Replacing all instances of S^c with D^c in (3.3) and (3.4), this implies that $D \sim n$.

To find $\mathbb{E}[D^c]$, we must find $P(k \notin A - A)$ for every $k \in \mathbb{Z}/n\mathbb{Z}$. First, we assume that $A \neq \emptyset$, because this happens with negligible probability since we are in the slow decay case. Because $A \neq \emptyset$, we only consider $k \neq 0$. Having fixed k, there are n different pairs (a, b) such that $a - b \equiv k \mod n$: $(k, 0), (2k, k), \dots, ((n-1)k, (n-2)k), (0, (n-1)k)$.

The pairs are all ordered because subtraction isn't commutative. Then $k \notin A - A$ if and only if

$$(0 \notin A \vee k \notin A) \ \wedge \ (k \notin A \vee 2k \notin A) \ \wedge \ \cdots \ \wedge \ ((n-1)k \notin A \vee 0 \notin A). \qquad (3.13)$$

Similarly to the previous section, this lends itself to a natural graph-theoretic interpretation. We construct the graph $G_{n,k}$ with vertex set $V = \{0, 1, \dots, n-1\}$ and with edge set $E = \{\{0, k\}, \dots, \{(n-1)k, 0\}\}$. In other words, we draw an edge between all vertices a and b such that $a - b \equiv k \mod n$ or $b - a \equiv k \mod n$. Then an equivalent formulation of (3.13) is that $k \notin A - A$ if and only if no two adjacent vertices of $G_{n,k}$ are in A.

Because we assume n is prime and $k \not\equiv 0 \pmod{n}$, all of $0, k, 2k, \ldots, (n-1)k$ are distinct mod n, so $G_{n,k}$ is necessarily a cycle on n vertices (see Figure 3 for an example).

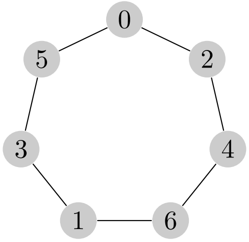

Figure 3: The graph $G_{7,2}$.

If we re-label each vertex ak as a, then picking an $A \subseteq \mathbb{Z}/n\mathbb{Z}$ such that $k \notin A - A$ is equivalent to picking a subset of $\{0, 1, \ldots, n-1\}$ such that no two consecutive elements are picked, where 0 and $n-1$ are considered to be consecutive. By the calculation of the quantity $D(n,k)$ in Theorem B.1 from Appendix B, there are $\binom{n-r+1}{r} - \binom{n-r-1}{r-2}$ ways to choose such an r-element subset of $\{0, 1, \ldots, n-1\}$. We have then that

$$P(k \notin A - A) \;=\; \sum_{r=1}^{\lfloor n/2 \rfloor} \left[\binom{n-r+1}{r} - \binom{n-r-1}{r-2}\right] p^r (1-p)^{n-r}. \tag{3.14}$$

We start the summation at $r = 1$ because we have assumed $A \neq \emptyset$. We sum until $r = \lfloor n/2 \rfloor$ because $\binom{n-r+1}{r} - \binom{n-r-1}{r-2}$ is zero for all bigger r.

Remark 3.4. Here is where we rely heavily on the assumption that n is prime. If n is not prime, then the graph $G_{n,k}$ becomes a union of disjoint cycles of length $n/\gcd(n,k)$, and so $\operatorname{Prob}(k \notin A - A)$ becomes

$$\left(\sum_{r=1}^{\lfloor n/(2d(k)) \rfloor} \left[\binom{n/d(k) - r + 1}{r} - \binom{n/d(k) - r - 1}{r-2}\right] p^r (1-p)^{n/d(k) - r}\right)^{d(k)}, \tag{3.15}$$

where $d(k) = \gcd(n,k)$. Simulations suggest that as $n \to \infty$, this quantity is independent of $d(k)$, but the analysis becomes significantly more involved.

We have then that

$$\begin{aligned}\mathbb{E}[D^c] &= \sum_{k\in(\mathbb{Z}/n\mathbb{Z})\setminus\{\bar{0}\}} P(x \notin A-A) \\ &= (n-1)\sum_{r=1}^{\lfloor n/2\rfloor}\left[\binom{n-r+1}{r}-\binom{n-r-1}{r-2}\right]p^r(1-p)^{n-r} \\ &\leq n\sum_{r=0}^{n/2}\left[\binom{n-r+1}{r}-\binom{n-r-1}{r-2}\right]p^r(1-p)^{n-r} \\ &= n\sum_{r=0}^{n/2}\left[\frac{(n-r-1)!(n^2-2nr+n)}{r!(n-2r+1)!}\right]p^r(1-p)^{n-r} \\ &= n\sum_{r=0}^{n/2}\left[\frac{n(n-r)!}{r!(n-2r)!(n-r)}\right]p^r(1-p)^{n-r} \\ &= n\sum_{r=0}^{n/2}\binom{n-r}{r}\frac{n}{n-r}p^r(1-p)^{n-r} \\ &\leq 2n\sum_{r=0}^{n/2}\binom{n-r}{r}p^r(1-p)^{n-r} \ = \ 2nF(n), \qquad (3.16)\end{aligned}$$

and this quantity tends to 0 by Lemma 3.1.

We compute $\mathrm{Var}(D^c)$ in a similar manner as $\mathrm{Var}(S^c)$. Define the random variables

$$Z'_k \ := \ \begin{cases}1, & k \notin A-A \\ 0, & k \in A-A.\end{cases} \qquad (3.17)$$

We have

$$\begin{aligned}\mathrm{Var}(D^c) &= \sum_{k\in\mathbb{Z}/n\mathbb{Z}}\mathrm{Var}(Z'_k)+\sum_{i\neq j\in\mathbb{Z}/n\mathbb{Z}}\mathrm{Cov}\big(Z'_i,Z'_j\big) \\ &\sim \sum_{k\neq 0}\big(\mathbb{E}[(Z'_k)^2]-\mathbb{E}[Z'_k]^2\big)+\sum_{i\neq j}\big(\mathbb{E}[Z'_iZ'_j]-\mathbb{E}[Z'_i]\mathbb{E}[Z'_j]\big) \\ &\sim nF(n)-n^2F(n)^2+\sum_{i\neq j}\mathrm{Prob}\,(i\notin A-A\wedge j\notin A-A)\,. \qquad (3.18)\end{aligned}$$

Now note that in particular, $\mathrm{Prob}\,(i\notin A-A\wedge j\notin A-A)\leq \mathrm{Prob}\,(i\notin A-A)$, so we have the bound

$$\begin{aligned}\mathrm{Var}(D^c) &\leq nF(n)-(nF(n))^2+n(n-1)F(n) \\ &\leq n^2F(n)-(nF(n))^2, \qquad (3.19)\end{aligned}$$

which tends to 0 by Lemma 3.1. This completes the proof of part (3) of Theorem 1.1. □

References

[1] N. Alon and J.H. Spencer, *The Probabilistic Method*, Wiley, 1992.

[2] T. Do, A. Kulkarni, S.J. Miller, D. Moon, J. Wellens, and J. Wilcox, Sets characterized by missing sums and differences in dilating polytopes, *J. Number Theory*, 157 (2015), 123-153.

[3] P. Hegarty and S.J. Miller, When almost all sets are difference dominated", *Random Structures and Algorithms*, **35** (2009), no. 1, 118-136.

[4] R. Honsberger, A Second Look at the Fibonacci and Lucas Numbers, Chapter 8 in *Mathematical Gems III*, Mathematical Association of America, Washington, DC, 1985.

[5] M. Koloğlu, G. Kopp, S. J. Miller and Y. Wang, On the number of summands in Zeckendorf decompositions, *Fibonacci Quart.* **49** (2011), no. 2, 116–130.

[6] T. Koshy, *Fibonacci and Lucas Numbers with Applications*, New York: Wiley, 2001.

[7] O. Lazarev, S. J. Miller, and K. O'Bryant, Distribution of Missing Sums in Sumsets, *Exp. Math.*, **22** (2013), no. 2, 132-156.

[8] G. Martin and K. O'Bryant, Many sets have more sums than differences, in *Additive Combinatorics*, CRM Proc. Lecture Notes, vol. 43, American Mathematical Society, Providence, RI, 2007, pp. 287–305.

[9] S. J. Miller and K. Vissuet, Most subsets are balanced in finite groups, *Combinatorial and Additive Number Theory: CANT 2011 and 2012 (Springer Proceedings in Mathematics & Statistics)* (2014), 147-157.

[10] M. Nathanson, Problems in Additive Number Theory, I, *Additive combinatorics*, CRM Proc. Lecture Notes **43**, Amer. Math. Soc., Providence, RI, 2007, pp. 263-270.

[11] Y. Zhao, Sets characterized by missing sums and differences, *J. Number Theory* **131** (2011), no. 11, 2107-2134.

Appendix

A. Proof of Lemma 3.1

Lemma 3.1. Suppose $p(n) = n^{-\delta}$ where $\delta \in (0, 1/2)$. Let

$$F(n) \;=\; \sum_{r=0}^{n/2} \binom{n-r}{r} p^r (1-p)^{n-r}. \tag{A.1}$$

Then $F(n) = o(1/n^3)$.

Proof. We use the following well-known approximations.

- Binomial approximation: if X and Y are two quantities depending on n where $1 = o(X)$ and $Y = o(X)$, then

$$\binom{X}{Y} \sim \frac{X^Y}{Y!}. \tag{A.2}$$

- Stirling's formula:

$$n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n \tag{A.3}$$

With these at our disposal, we can prove Lemma 3.1. Note that for any r,

$$\binom{n-r}{r}p^r(1-p)^{n-r} \;\leq\; \binom{n}{r}p^r(1-p)^{n-r}. \tag{A.4}$$

Since the binomial distribution has mean np and variance $np(1-p)$, we have by Chebyshev's inequality

$$\begin{aligned} &\sum_{|r-np|\;\geq\;\log n\sqrt{np(1-p)}} \binom{n}{r}p^r(1-p)^{n-r} \\ &\qquad = \; o\left(\sum_{|r-np|\;<\;\log n\sqrt{np(1-p)}} \binom{n}{r}p^r(1-p)^{n-r}\right). \end{aligned} \tag{A.5}$$

What this is saying is that the tails of the distribution are negligible compared to the middle. Thus, by (A.4), we can write

$$F(n) \;=\; \sum_{r=0}^{n/2}\binom{n-r}{r}p^r(1-p)^{n-r} \;\sim\; \sum_{|r-np|\;<\;\log n\sqrt{np(1-p)}}\binom{n-r}{r}p^r(1-p)^{n-r}. \tag{A.6}$$

What this means is that all but a negligible amount of the contribution to $F(n)$ comes from the terms where r is close to np.

We have

$$\begin{aligned} \binom{n-r}{r}p^r(1-p)^{n-r} \;&\leq\; (1-p)^n\frac{(n-r)^r}{r!}\left(\frac{p}{1-p}\right)^r && \text{(A.7)} \\ &\leq\; 2(1-p)^n\frac{(n-r)^r}{\sqrt{2\pi r}(r/e)^r}\left(\frac{p}{1-p}\right)^r && \text{(A.8)} \\ &\leq\; 2(1-p)^n\left(\frac{(n-r)ep}{r(1-p)}\right)^r. && \text{(A.9)} \end{aligned}$$

The inequality in (A.7) comes from the binomial approximation and the inequality in (A.8) comes from Stirling's formula. Denote the quantity on the right side of

(A.9) by $g(r)$. We now maximize $\log g(r)$ over $r \in [0, n/2]$.

$$\frac{g'(r)}{g(r)} \;=\; \log\left(\frac{(n-r)pe}{r(1-p)}\right) - \frac{n}{n-r}. \tag{A.10}$$

It is clear that $g(r)$ is small at the endpoints; thus $g(r)$ is maximized at $r = r_0$ such that

$$\log\left(\frac{(n-r_0)ep}{r_0(1-p)}\right) \;=\; \frac{n}{n-r_0}. \tag{A.11}$$

We know by (A.6) that r_0 must satisfy $|r_0 - np| \;<\; \log n\sqrt{np(1-p)}$, so we have

$$\log\left(\frac{(n-r_0)ep}{r_0(1-p)}\right) \;\sim\; 1; \tag{A.12}$$

thus $r_0 \sim np$. Letting $r = np$ in (A.9), we now have the bound

$$\begin{aligned}\binom{n-r}{r}p^r(1-p)^{n-r} \;&\le\; 2(1-p)^n\left(\frac{(n-np)pe}{np(1-p)}\right)^{np}\\ &\le\; 2(1-p)^n e^{np}\\ &=\; 2(e^p - pe^p)^n, \end{aligned} \tag{A.13}$$

so that

$$F(n) \;\le\; 2n(e^p - pe^p)^n. \tag{A.14}$$

To complete the proof, it suffices to show that $h(n) := 2n^4(e^p - pe^p)^n \to 0$ as $n \to \infty$. We have

$$\begin{aligned}\log h(n) \;&=\; \log 2 + 4\log n + n\log(e^p(1-p))\\ &=\; \log n + n\log(1-p) + np\\ &=\; \log n + n(-p - \frac{1}{2}p^2 + O(p^3)) + np\\ &=\; \log n - \frac{1}{2}np^2 + O(np^3)\end{aligned} \tag{A.15}$$

and this tends to $-\infty$ as $n \to \infty$ because $\log n = o(np^2)$; thus $h(n)$ tends to 0. This completes the proof of Lemma 3.1. □

Remark A.1. As in Remark 3.2, if $p(n) = o(\sqrt{\log n} \cdot n^{-1/2})$, then $\log h(n)$ tends to $+\infty$ rather than $-\infty$.

B. Note on Lucas Numbers

The Lucas numbers are defined by the recurrence

$$L_n = L_{n-1} + L_{n-2}$$

with initial values $L_0 = 2$ and $L_1 = 1$. Combinatorially, the n-th Lucas number represents the number of subsets of $\{1, \ldots, n\}$ containing no consecutive integers, where 1 and n are counted as consecutive (see [4] for a proof of this). An equivalent formulation of the following formula appears on page 173 of [6], but we use a different counting argument to establish it directly. We prove the following:

Theorem B.1. *For all $n \geq 2$,*

$$L_n = \sum_{k=0}^{\lfloor n/2 \rfloor} \left[\binom{n-k+1}{k} - \binom{n-k-1}{k-2} \right].$$

Proof. Let $D(n,k)$ denote the number of k-element subsets of $\{1, \ldots, n\}$ containing no two consecutive integers, where 1 and n are considered consecutive. Note that for any $k > n/2$, the pigeonhole principle forces $D(n,k) = 0$. Thus

$$\sum_{k=0}^{\lfloor n/2 \rfloor} D(n,k) = L_n, \tag{B.1}$$

and we just need to show $D(n,k) = \binom{n-k+1}{k} - \binom{n-k-1}{k-2}$. For fixed n, k, let

$Y = \#$ acceptable subsets without considering $1, n$ consecutive

$Z = \#$ subsets that contain both 1 and n but no other consecutive integers

and note that $D(n,k) = Y - Z$. Note also that $Y = C(n,k)$ from (3.9).

To count Y, we use a standard stars-and-bars argument. Suppose you have n objects in a row, and you need to select k of them, no two of which are consecutive. Remove k of the objects. You now need to reinsert the k objects into the row such that no two are consecutive, which means you have $n - k + 1$ spots to choose from (one spot in between each remaining pair of objects and one on each end of the row). Thus the number of ways to pick k non-consecutive elements from a row of n is $\binom{n-k+1}{k}$.

Now note that to count Z, we just repeat the argument for Y, but this time we are picking $k-2$ non-consecutive elements from $\{3, \ldots, n-2\}$, and there are $\binom{(n-4)-(k-2)+1}{k-2} = \binom{n-k-1}{k-2}$. So $D(n,k) = \binom{n-k+1}{k} - \binom{n-k-1}{k-2}$. □