involve
a journal of mathematics

msp

2019 vol. 12, no. 7

involve

msp.org/involve

INVOLVE YOUR STUDENTS IN RESEARCH

Involve showcases and encourages high-quality mathematical research involving students from all academic levels. The editorial board consists of mathematical scientists committed to nurturing student participation in research. Bridging the gap between the extremes of purely undergraduate research journals and mainstream research journals, *Involve* provides a venue to mathematicians wishing to encourage the creative involvement of students.

See inside back cover or msp.org/involve for submission instructions. The subscription price for 2019 is US $195/year for the electronic version, and $260/year (+$35, if shipping outside the US) for print and electronic. Subscriptions, requests for back issues and changes of subscriber address should be sent to MSP.

Involve (ISSN 1944-4184 electronic, 1944-4176 printed) at Mathematical Sciences Publishers, 798 Evans Hall #3840, c/o University of California, Berkeley, CA 94720-3840, is published continuously online. Periodical rate postage paid at Berkeley, CA 94704, and additional mailing offices.

Involve peer review and production are managed by EditFlow® from Mathematical Sciences Publishers.

PUBLISHED BY

mathematical sciences publishers

nonprofit scientific publishing

http://msp.org/

INVOLVE 12:7 (2019)
dx.doi.org/10.2140/involve.2019.12.1081

Asymptotic expansion of Warlimont functions on Wright semigroups

Marco Aldi and Hanqiu Tan

(Communicated by Kenneth S. Berenhaut)

We calculate full asymptotic expansions of prime-independent multiplicative functions on additive arithmetic semigroups that satisfy a strong form of Knopfmacher's axioms. When applied to the semigroup of unlabeled graphs, our method yields detailed asymptotic information on how graphs decompose into connected components. As a second class of examples, we discuss polynomials in several variables over a finite field.

1. Introduction

Let G_n be the number of unlabeled graphs with n vertices and let G_n^+ be the number of connected unlabeled graphs with n vertices. Using the fact that the sequences $\{G_n\}$ and $\{G_n^+\}$ are related by the identity

$$\sum_{n=0}^{\infty} G_n x^n = \prod_{m=1}^{\infty} (1-x^m)^{-G_m^+}, \tag{1}$$

Wright [1967] proved that $G_n \sim G_n^+$; i.e., almost all graphs are connected. As observed in that paper and further clarified in [Warlimont 2001], this asymptotic result is intimately related to the fact that the power series at the right-hand side of (1) has trivial convergence radius. Armed with a full asymptotic expansion for G_n [Wright 1969], Wright [1970] further improved this result by constructing a sequence $\{\omega_s\}$ of polynomials such that, for any fixed positive integer R, the asymptotic relation

$$G_n^+ = G_n + \sum_{s=1}^{R-1} \omega_s(n) G_{n-s} + O(n^R G_{n-R}) \tag{2}$$

holds in the limit $n \to \infty$.

In the context of abstract analytic number theory [Knopfmacher 1975], Knopfmacher [1976] (see also [Flajolet and Sedgewick 2009; Burris 2001] for the more

MSC2010: primary 05A16; secondary 05C30, 11T06.
Keywords: arithmetical semigroups, asymptotic enumeration, graph enumeration.

general setting of weighted decomposable combinatorial structures) observed that (1) is a particular case of an Euler product type of identity which holds for arbitrary additive arithmetical semigroups and that the methods of [Wright 1967] can be used to study the distribution of certain arithmetical functions on additive arithmetical semigroups in which almost all elements are prime. For instance, if d_2 is the *divisor function* that to each unlabeled graph g assigns the number of ways to write g as a disjoint union of an ordered pair of graphs then

$$\lim_{n\to\infty} \frac{1}{G_n} \sum d_2(g) = 2 \quad \text{and} \quad \lim_{n\to\infty} \frac{1}{G_n} \sum (d_2(g) - 2)^2 = 0, \tag{3}$$

where both sums are taken over all graphs g with n vertices.

The goal of the present paper is to investigate Knopfmacher's suggestion [1976] that restricting to arithmetical semigroups in which the total number of elements is related to the number of prime elements by a formula analogous to (2) might lead to a strengthening of (3). To illustrate our results with an example, consider again the divisor function d_2 on the semigroup of graphs. We prove that for every positive integer M, there exists a sequence $\{\tau_s(n)\}$ of polynomials such that, for any fixed positive integer R, the asymptotic relation

$$\frac{1}{G_n} \sum (d_2(g) - 2)^M = 2^M \sum_{s=1}^{R-1} \tau_s(n) 2^{-sn} + O(n^{2R-1} 2^{-Rn}) \tag{4}$$

holds in the limit $n \to \infty$. Clearly, (3) can be recovered by setting $M = 1$ and $M = 2$ in (4) and taking the limit as $n \to \infty$. More generally, we show that (4) is a particular case of a formula that holds if d_2 is replaced by an arbitrary *Warlimont function*, i.e., a multiplicative prime-independent function whose restriction to power of primes grows in a prescribed way. Even more generally, the semigroup of graphs can be replaced by any *Wright semigroup*, which we define to be an additive arithmetical semigroup subject to a growth condition introduced in [Wright 1970]. Examples of Wright semigroups include the semigroup of unlabeled graphs with an even number of edges and the semigroup of polynomials in at least two variables over a finite field.

The paper is organized as follows. Section 2 contains the main technical results used in the rest of the paper. We work with triples of sequences related by a generalization of (1) that were introduced in [Warlimont 1993]. The main result is Theorem 5 which can be thought of as a generalization of [Wright 1970], modeled after the way in which [Warlimont 1993] generalizes [Wright 1967]. In Section 3, after introducing the key notions of Wright semigroup and of Warlimont function, we provide asymptotic formulas for moments of Warlimont functions in terms of the number of elements of given degree in the underlying (not necessarily Wright) semigroup. In the special case of Wright semigroup, we construct full

asymptotic expansions generalizing (4). We illustrate our results in Section 4 by calculating asymptotic expansion of some of the arithmetical functions considered in [Knopfmacher 1976] on three examples of Wright semigroups: the semigroup of all unlabeled graphs, the semigroup of unlabeled graphs with an even number of edges and the semigroup of nonzero polynomials (up to scaling) in at least two variables over a finite field.

2. Warlimont triples

Definition 1. A *Warlimont triple* is a triple $(\{T_n\}, \{t_n\}, \{a_n\})$ of sequences of nonnegative real numbers related by the identity

$$\sum_{n=0}^{\infty} T_n x^n = \prod_{m=1}^{\infty} \Big(\sum_{k=0}^{\infty} a_k x^{km} \Big)^{t_m} \tag{5}$$

of formal power series and such that

(i) $T_0 = a_0 = 1$,

(ii) $a_1 > 0$,

(iii) $t_m \in \mathbb{Z}$ for all m and $t_m > 0$ for all but finitely many m.

In order for the three sequences to be all indexed by nonnegative integers, we set $t_0 = 0$.

Lemma 2. *Let $(\{T_n\}, \{t_n\}, \{a_n\})$ be a Warlimont triple and consider the sequences $\{v_n\}$, $\{\beta_n\}$ and $\{b_n\}$ defined the recursion formulas*

$$v_n = T_n - \sum_{s=1}^{n-1} \frac{s}{n} v_s T_{n-s}, \tag{6}$$

$$\beta_n = -\sum_{s=0}^{n-1} \beta_s T_{n-s}, \tag{7}$$

$$b_n = n a_n - \sum_{s=1}^{n-1} b_s a_{n-s}, \tag{8}$$

with initial conditions $v_1 = T_1$, $\beta_0 = 1$, $b_1 = a_1$. Then for all n:

(i) $v_n = \sum_{d \mid n} (d/n) t_d b_{n/d}$, *where the sum is over all integers $1 < d \le n$ that divide n.*

(ii) $\beta_n = -\sum_{s=1}^{n} (s/n) v_s \beta_{n-s}$.

(iii) *For every positive integer R*

$$\sum_{s=0}^{R-1} \beta_s T_{n-s} = v_n + \frac{1}{n} \sum_{r=0}^{R-1} \beta_r \sum_{s=R-r}^{n-R} s v_s T_{n-r-s}.$$

Proof. Using formal term-by-term differentiation it is easy to show that (6) and (7) are equivalent to the formal identities

$$\log\left(\sum_{n=0}^{\infty} T_n x^n\right) = \sum_{m=1}^{\infty} v_m x^m, \tag{9}$$

$$\log\left(\sum_{n=0}^{\infty} a_n x^n\right) = \sum_{s=1}^{\infty} \frac{b_s}{s} x^s, \tag{10}$$

respectively. Taking the formal logarithm of (5) and substituting (9), (10), we obtain

$$\sum_{m=1}^{\infty} v_m x^m = \sum_{r,s=1}^{\infty} t_r \frac{b_s}{s} x^{rs}$$

from which (i) easily follows. Since (7) is equivalent to the identity

$$\left(\sum_{s=0}^{\infty} \beta_s x^s\right)\left(\sum_{n=0}^{\infty} T_n x^n\right) = 1$$

of formal power series, taking formal logarithms yields

$$\log\left(\sum_{s=0}^{\infty} \beta_s x^s\right) = -\sum_{m=1}^{\infty} v_m x^m.$$

Comparing with (6) and (9) proves (ii). It follows from (ii) that

$$\sum_{u=0}^{R-1}\sum_{r=0}^{u} \beta_r((n-u)v_{n-u}T_{u-r} + (u-r)v_{u-r}T_{n-u}) = v_n - \sum_{u=0}^{R-1} u\beta_u T_{n-u}$$

and thus

$$\begin{aligned}
&\sum_{s=0}^{R-1} \beta_s T_{n-s} - v_n \\
&\quad = \frac{1}{n}\sum_{u=0}^{R-1}\sum_{r=0}^{u} \beta_r\left(\frac{n-r}{R-r}T_{n-r} - (n-u)v_{n-u}T_{u-r} - (u-r)v_{u-r}T_{n-u}\right) \\
&\quad = \frac{1}{n}\sum_{r=0}^{R-1} \beta_r\left((n-r)T_{n-r} - \sum_{s=0}^{R-r-1}(n-r-s)T_s v_{n-r-s} + \sum_{s=0}^{R-r-1} s v_s T_{n-r-s}\right) \\
&\quad = \frac{1}{n}\sum_{r=0}^{R-1} \beta_r \sum_{s=R-r}^{n-R} s v_s T_{n-r-s},
\end{aligned}$$

where the last line follows from applying (6) to T_{n-r} for each $r \in \{0, 1, \ldots, R-1\}$. □

Lemma 3. *Let* $(\{T_n\}, \{t_n\}, \{a_n\})$ *be a Warlimont triple. Then*

$$a_1 \sum_{s=0}^{\lfloor n/2 \rfloor} T_s t_{n-s} \le T_n$$

for all n.

Proof. Since

$$\prod_{m=N+1}^{\infty} \left(\sum_{k=0}^{\infty} a_k x^{km} \right)^{t_m} \in 1 + x^{N+1} \mathbb{R}[\![x]\!]$$

for every integer $N \ge 0$, we have

$$\sum_{n=0}^{N} T_n x^n \in \prod_{m=1}^{N} \left(\sum_{k=0}^{\infty} a_k x^{km} \right)^{t_m} + x^{N+1} \mathbb{R}[\![x]\!]$$

and thus

$$\sum_{n=0}^{N} T_n x^n \in \left(\sum_{s=0}^{\lfloor N/2 \rfloor} T_s x^s \right) \prod_{m=\lfloor N/2 \rfloor + 1}^{N} \left(\sum_{k=0}^{\infty} a_k x^{km} \right)^{t_m} + x^{N+1} \mathbb{R}[\![x]\!]. \tag{11}$$

On the other hand, by assumption t_m is a nonnegative integer for all m and thus by the binomial theorem

$$\left(\sum_{k=0}^{\infty} a_k x^{km} \right)^{t_m} \in 1 + a_1 t_m x^m + x^{2m} \mathbb{R}[\![x]\!]. \tag{12}$$

Since the sequences $\{a_k\}$, $\{t_m\}$ and $\{T_n\}$ are nonnegative, the lemma follows by substituting (12) into (11) and comparing coefficients. □

Lemma 4. *Let* $(\{T_n\}, \{t_n\}, \{a_n\})$ *be a Warlimont triple such that* $\log(a_n) = O(n)$. *Then for every nonnegative integer* R

$$|v_n - a_1 t_n| = \begin{cases} O(T_{n-R}) & \text{if } T_{n-1} = o(T_n), \\ O(t_{n-R}) & \text{if } t_{n-1} = o(t_n). \end{cases}$$

Proof. Assume $T_{n-1} = o(T_n)$. Since $\log(a_n) = O(n)$, there exists $r > 1$ such that $a_n \le r^n$ for all n. By induction on the definition of $\{b_n\}$, we obtain $|b_n| \le (3r)^n$ for all n. Moreover, since $T_{n-1} = o(T_n)$ and condition (iii) in the definition of Warlimont triple implies $T_n > 0$ for all but finitely many n, there exists a positive integer N such that $0 < T_n \le (3r)^{-2} T_{n+1}$ for all $n \ge N$. If

$$C = \max\left\{ 1, \frac{(3r)^{2N} T_0}{T_N}, \frac{(3r)^{2N} T_1}{T_{N+1}}, \dots, \frac{(3r)^{2N} T_{N-1}}{T_{2N-1}} \right\}$$

then for any $n > 0$ and for any $m \ge N$ we obtain

$$T_n \le C(3r)^{-2N} T_{n+N} \le C(3r)^{-2(N+1)} T_{n+N+1} \le \cdots \le C(3r)^{-2m} T_{n+m}.$$

Therefore, using Lemmas 2 and 3

$$|v_n - a_1 t_n| \le \sum_{d/n} \frac{d}{n} t_d |b_{n/d}| \le \sum_{d/n} T_d (3r)^{n/d} \le C T_{n-R} \sum_{d/n} (3r)^{-n+2R+2d} = O(T_{n-R}).$$

The proof for the case $t_{n-1} = o(t_n)$ is similar and left to the reader. □

Theorem 5. *Let* $(\{T_n\}, \{t_n\}, \{a_n\})$ *be a Warlimont triple such that* $\log(a_n) = O(n)$ *and let* R *be a fixed positive integer. Then the following are equivalent*:

(i) $T_{n-1} = o(T_n)$ *and*

$$\sum_{s=R}^{n-R} T_s T_{n-s} = O(T_{n-R}).$$

(ii) $T_{n-1} = o(T_n)$ *and*

$$a_1 t_n = \sum_{s=0}^{R-1} \beta_s T_{n-s} + O(T_{n-R}).$$

(iii) $t_{n-1} = o(t_n)$ *and*

$$T_n = a_1 \sum_{s=0}^{R-1} T_s t_{n-s} + O(t_{n-R}).$$

(iv) $t_{n-1} = o(t_n)$ *and*

$$\sum_{s=R}^{n-R} t_s t_{n-s} = O(t_{n-R}).$$

Proof. Assume (i) holds. Using Lemmas 3 and 4 and $T_{n-1} = o(T_n)$, we obtain

$$|v_n| \le |v_n - a_1 t_n| + a_1 t_n = O(T_n).$$

Therefore, there exist an integer $N > R$ and a constant $C > 0$ such that $|v_n| \le C T_n \le C T_{n+r}$ for all $n \ge N$ and for all $r \in \{0, \dots, R-1\}$. Combining this observation with Lemma 2 yields

$$\begin{aligned}
\left| v_n - \sum_{s=0}^{R-1} \beta_s T_{n-s} \right| &\le \sum_{r=0}^{R-1} \beta_r \sum_{s=R-r}^{n-R} \frac{s}{n} |v_s| T_{n-s} \\
&\le \sum_{r=0}^{R-1} \beta_r \left(\sum_{s=R-r}^{N-1} |v_s| T_{n-r-s} + C \sum_{s=N}^{n-R} T_s T_{n-r-s} \right) \\
&\le \sum_{r=0}^{R-1} \beta_r \left(\sum_{s=R-r}^{N-1} |v_s| T_{n-r-s} + C^2 \sum_{s=R}^{n-R} T_s T_{n-s} \right) \\
&= O(T_{n-R}),
\end{aligned}$$

and thus

$$\left|a_1 t_n - \sum_{s=0}^{R-1} \beta_s T_{n-s}\right| \le |a_1 t_n - v_n| + \left|v_n - \sum_{s=0}^{R-1} \beta_s T_{n-s}\right| = O(T_{n-R}).$$

Hence, (i) implies (ii). Assume (ii) holds. Then

$$\begin{aligned} a_1 \sum_{s=0}^{R-1} T_s t_{n-s} &= \sum_{s=0}^{R-1} T_s \sum_{r=0}^{R-1-s} \beta_r T_{n-s-r} + O(T_{n-R}) \\ &= \sum_{s=0}^{R-1} \sum_{u=s}^{R-1} T_s \beta_{u-s} T_{n-u} + O(T_{n-R}) \\ &= \sum_{u=0}^{R-1} T_{n-u} \sum_{s=0}^{u} T_s \beta_{u-s} + O(T_{n-R}) \\ &= T_n + O(T_{n-R}), \end{aligned}$$

where the last equality is obtained using the definition of the sequence $\{\beta_n\}$. In particular, setting $R = 1$ we obtain $a_1 t_n - T_n = O(T_{n-1}) = o(T_n)$, which implies $a_1 t_n \sim T_n$ and hence $o(t_{n-1}) = O(T_{n-1}) = o(T_n) = o(t_n)$. Therefore, (ii) implies (iii). Assume (iii) holds. By Lemma 3

$$a_1^2 \sum_{s=R}^{\lfloor n/2 \rfloor} t_s t_{n-s} \le a_1 \sum_{s=R}^{\lfloor n/2 \rfloor} T_s t_{n-s} \le T_n - a_1 \sum_{s=0}^{R-1} T_s t_{n-s} = O(t_{n-R}).$$

This proves (iv) since

$$\sum_{s=R}^{n-R} t_s t_{n-s} = 2 \sum_{s=R}^{\lfloor n/2 \rfloor} t_s t_{n-s} + O(t_{n-R}).$$

Finally, assume (iv) holds. Lemma 4 implies $|v_n - a_1 t_n| = O(t_{n-R}) = o(t_n)$ and thus $v_n \sim a_1 t_n$. This implies that there exist an integer $N \ge R$ and constants $c, C > 0$ such that

$$0 < c t_n \le v_n \le C t_n \tag{13}$$

for all $n \ge N$. As a consequence,

$$v_{n-1} = O(t_{n-1}) = o(t_n) = o(v_n) \tag{14}$$

and

$$\sum_{j=N}^{n-N} v_{n-j} v_j \le C^2 \sum_{j=R}^{n-R} t_{n-j} t_j = O(t_{n-R}) = O(v_{n-R}). \tag{15}$$

For each $n \ge N$, let

$$M_n = \max\left\{ \frac{T_j}{v_j} \;\middle|\; N \le j \le n \right\}.$$

By Lemma 2, we obtain

$$|T_n - v_n| \le \sum_{j=1}^{n-1} |v_{n-j}|T_j \le \sum_{j=1}^{N-1} v_{n-j}T_j + M_{n-1}\left(\sum_{j=N}^{n-N} v_{n-j}v_j + \sum_{j=n-N+1}^{n-1} |v_{n-j}|v_j\right)$$
$$= o(v_n)(1+M_{n-1}),$$

where (14) and (15) were used to obtain the last equality. Hence there exists $N_1 \ge N$ such that for all $n \ge N_1$

$$\frac{T_n}{v_n} \le \frac{3+M_{n-1}}{2}$$

and thus

$$M_n = \max\left\{M_{n-1}, \frac{T_n}{v_n}\right\} \le \max\{M_{n-1}, 3\}.$$

This shows that the sequence $\{M_n\}$ is bounded; i.e., there exists a constant $K > 0$ such that $T_n \le K v_n$ for all $n \ge N$. Therefore, using (13) and Lemma 3, we obtain

$$T_{n-1} = O(v_{n-1}) = O(t_{n-1}) = o(t_n) = o(T_n).$$

Moreover (14) yields

$$\sum_{s=R}^{n-R} T_{n-s}T_s \le 2\sum_{s=R}^{N} T_s T_{n-s} + K^2 \sum_{s=N}^{n-N} v_{n-s}v_s = O(T_{n-R}) + O(v_{n-R}) = O(T_{n-R}).$$

This concludes the proof that (iv) implies (i) and the theorem is proved. □

Remark 6. Let $(\{T_n\}, \{t_n\}, \{a_n\})$ be a Warlimont triple that satisfies the equivalent conditions of Theorem 5 for some $R > 2$. Then

$$\sum_{s=R-1}^{n-R+1} T_s T_{n-s} = \sum_{s=R}^{n-R} T_s T_{n-s} + 2T_{R-1}T_{n-R+1} = O(T_{n-R}) + O(T_{n-R+1}) = O(T_{n-R+1})$$

and thus $(\{T_n\}, \{t_n\}, \{a_n\})$ satisfies the equivalent conditions of Theorem 5 for any fixed positive integer less than or equal to R. In particular, $t_{n-1} = o(t_n)$ and $T_n \sim a_1 t_n$.

3. Warlimont functions and Wright semigroups

Definition 7. An *additive arithmetical semigroup* is a pair $(G, +, \partial)$ consisting of an abelian semigroup $(G, +)$ with identity and a semigroup homomorphism $\partial : (G, +) \to (\mathbb{Z}_{\ge 0}, +)$ such that

(i) the cardinality G_n of the preimage $\partial^{-1}(n)$ is finite for all n,

(ii) G is freely generated by $G^+ \subseteq G$.

We denote by G_n^+ the cardinality of the set $\partial^{-1}(n) \cap G^+$.

Remark 8. Let $(G, +, \partial)$ be an additive arithmetical semigroup. As pointed out in [Knopfmacher 1976; Warlimont 1993], $(\{G_n\}, \{G_n^+\}, \{1\})$ is a Warlimont triple.

Definition 9. A *Wright semigroup* is an additive arithmetical semigroup $(G, +, \partial)$ satisfying

$$\log(G_n) = \alpha n^{a+1} + \beta n \log(n) + \gamma n + O(n^b) \tag{16}$$

for some real numbers α, β, γ, a, b such that $\alpha > 0$ and $0 < b < a$.

Definition 10. Let R be a positive integer. We say that an additive arithmetical semigroup $(G, +, \partial)$ satisfies *axiom* $\mathcal{W}_R$ if $G_{n-1} = o(G_n)$ and

$$\sum_{s=R}^{n-R} G_s G_{n-s} = O(G_{n-R}).$$

Remark 11. Let $(G, +, \partial)$ be an additive arithmetical semigroup that satisfies axiom $\mathcal{W}_R$ for some positive integer R. Combining Remarks 6 and 8, we conclude that $(G, +, \partial)$ satisfies axiom $\mathcal{W}_{R'}$ for any positive integer $R' \le R$. In particular, $G_n \sim G_n^+$ and $G_{n-1}^+ = o(G_n^+)$; i.e., the additive arithmetical semigroup $(G, +, \partial)$ satisfies both axiom $\mathcal{G}_1$ and axiom $\mathcal{G}_2$ as defined in [Knopfmacher 1976]. Notice that the combination of Axioms $\mathcal{G}_1$ and $\mathcal{G}_2$ is slightly weaker than axiom $\mathcal{W}_1$ since $\sum_{s=1}^{n-1} G_s G_{n-s} = o(G_n)$ does not necessarily imply $\sum_{s=1}^{n-1} G_s G_{n-s} = O(G_{n-1})$.

Proposition 12. *Every Wright semigroup satisfies axiom $\mathcal{W}_R$ for every positive integer R.*

Proof. This is a straightforward consequence of the definitions and Theorem 7 of [Wright 1970]. □

Definition 13. Let $(G, +, \partial)$ be an additive arithmetical semigroup. A function $F : G \to \mathbb{R}$ is *multiplicative* if $F(g_1 + g_2) = F(g_1)F(g_2)$ for all $g_1, g_2 \in G$ coprime. We say that F is *prime-independent* if there exists a sequence $\{F_n^+\}$ such that $F_n^+ = F(np)$ for every $p \in G^+$ and every positive integer n. For every function $F : G \to \mathbb{R}$, we denote by $\{F_n\}$ the sequence defined by setting

$$F_n = \sum_{\partial(g)=n} F(g)$$

for each nonnegative integer n. A *Warlimont function* is a nonnegative multiplicative prime-independent function such that $\log(F_n^+) = O(n)$ and $F_1^+ > 0$. The *normalization* of a Warlimont function F is the (not necessarily multiplicative) function $\widetilde{F} : G \to \mathbb{R}$ such that $\widetilde{F}(g) = F(g)/F_1^+$ for all $g \in G$.

Example 14. Let $(G, +, \partial)$ be an additive arithmetical semigroup and let $F : G \to \mathbb{R}$ be such that $F(g) = 1$ for all $g \in G$. Then F is a Warlimont function and $F_n = \widetilde{F}_n = G_n$ for all n.

Example 15. Let $(G, +, \partial)$ be an additive arithmetical semigroup and, for each $k \geq 2$, consider the *generalized divisor function* $d_k : G \to \mathbb{R}$ that to each $g \in G$ assigns the number $d_k(g)$ of k-tuples $(g_1, \ldots, g_k) \in G^k$ such that $g = g_1 + \cdots + g_k$. Then d_k is multiplicative, prime-independent and $(d_k)_n^+ = \binom{n+k-1}{k-1}$ for each integer $n \geq 1$. Therefore, d_k is Warlimont.

Example 16. Let $(G, +, \partial)$ be an additive arithmetical semigroup and consider the *unitary divisor function* $d_* : G \to \mathbb{R}$ that to each $g \in G$ assigns the number $d_*(g)$ of coprime pairs (g_1, g_2) such that $g = g_1 + g_2$. Then d_* is multiplicative, prime-independent and $(d_*)_n^+ = 2$ for each integer $n \geq 1$. Therefore d_* is Warlimont.

Example 17. Let $(G, +, \partial)$ be an additive arithmetical semigroup and consider the *prime divisor function* $B : G \to \mathbb{R}$ such that $B(k_1 p_1 + k_2 p_2 + \cdots + k_r p_r) = k_1 k_2 \cdots k_r$ for any $p_1, \ldots, p_r \in G$ primes and $k_1, \ldots, k_r$ positive integers. Then B is multiplicative, prime-independent and $B_n^+ = n$ for each integer $n \geq 1$. Therefore, B is Warlimont.

Remark 18. Let F be a Warlimont function on an additive arithmetical semigroup $(G, +, \partial)$. Then the function $F^m : G \to \mathbb{R}$ such that $F^m(g) = (F(g))^m$ for all $g \in G$ is again a Warlimont function for every integer $m \geq 1$ since

$$\log((F^m)_n^+) = m \log(F_n^+) = O(n).$$

Moreover, $\widetilde{F^m} = (\widetilde{F})^m$.

Remark 19. Let F be a Warlimont function on an additive arithmetical semigroup $(G, +, \partial)$. Then, as observed in [Warlimont 1993], $(\{F_n\}, \{G_n^+\}, \{F_n^+\})$ is a Warlimont triple.

Theorem 20. *Let $(G, +, \partial)$ be an additive arithmetical semigroup that satisfies axiom $\mathcal{W}_R$ and let F be a Warlimont function on G. Then for every positive integer M there exist constants $\xi_1, \ldots, \xi_{R-1}$ such that*

$$\sum_{\partial(g)=n} (\widetilde{F}(g) - 1)^M = \sum_{s=1}^{R-1} \xi_s G_{n-s} + O(G_{n-R}). \tag{17}$$

Proof. By Remark 19 and Example 14, $(\{G_n\}, \{G_n^+\}, \{1\})$ and $(\{F_n\}, \{G_n^+\}, \{F_n^+\})$ are both Warlimont triples. Since $\{G_n\}$ satisfies axiom $\mathcal{W}_R$, it follows from Theorem 5 applied to the Warlimont triple $(\{G_n\}, \{G_n^+\}, \{1\})$ that $G_{n-1}^+ = o(G_n^+)$,

$$\sum_{s=R}^{n-R} G_s^+ G_{n-s}^+ = O(G_{n-R}^+). \tag{18}$$

Moreover, if $\{\beta_n\}$ is the sequence defined recursively by setting $\beta_0 = 1$ and

$$\beta_n = -\sum_{s=0}^{n-1} \beta_s G_{n-s} \tag{19}$$

for every positive integer n, then

$$G^+_{n-s} = \sum_{r=0}^{R-1} \beta_r G_{n-s-r} + O(G_{n-s-R}) \tag{20}$$

for all $s \geq 0$. In particular, we can apply Theorem 5 to the Warlimont triple $(\{F_n\}, \{G^+_n\}, \{F^+_n\})$ and obtain

$$F_n = F^+_1 \sum_{s=0}^{R-1} F_s G^+_{n-s} + O(G^+_{n-R}). \tag{21}$$

Since by definition $G^+_n \leq G_n$ for all n and $G_{n-s-R} = o(G_{n-R})$ for all $s > 0$, substituting (20) into (21) yields

$$\widetilde{F}_n = \sum_{s=0}^{R-1} \left(\sum_{r=0}^{s} \beta_r F_{s-r} \right) G_{n-s} + O(G_{n-R}). \tag{22}$$

Using the binomial theorem and Remark 18 we obtain

$$\begin{aligned} \sum_{\partial(g)=n} (\widetilde{F}(g) - 1)^M &= (-1)^M \sum_{\partial(g)=n} \sum_{m=0}^{M} (-1)^m \binom{M}{m} \widetilde{F^m}(g) \\ &= (-1)^M \sum_{m=0}^{M} (-1)^m \binom{M}{m} (\widetilde{F^m})_n. \end{aligned} \tag{23}$$

Applying (22) to the Warlimont function F^m and substituting into the last line of (23) (after an obvious rearrangement) yields

$$\sum_{\partial(g)=n} (\widetilde{F}(g) - 1)^M = \sum_{s=0}^{R-1} \xi_s G_{n-s} + O(G_{n-R}), \tag{24}$$

with

$$\begin{aligned} \xi_s &= (-1)^M \sum_{m=0}^{M} (-1)^m \binom{M}{m} \sum_{r=0}^{s} \beta_r (F^m)_{s-r} \\ &= (-1)^M \sum_{m=1}^{M} (-1)^m \binom{M}{m} \sum_{r=0}^{s} \beta_r (F^m)_{s-r} \end{aligned} \tag{25}$$

for all $s \in \{0, \ldots, R-1\}$ where the second equality follows from (19) and Example 14. This implies (17) since, combining Remarks 18 and 19, $(F^m)_0 = 1$

for all m and thus

$$\xi_0 = (-1)^M \sum_{m=0}^{M} (-1)^m \binom{M}{m} \beta_0 (F^m)_0 = 0. \qquad \square$$

Definition 21. Let F be a Warlimont function on an additive arithmetical semigroup $(G, +, \partial)$ and let M be a positive integer. We define the *normalized M-th moments of F* to be the functions $\mu_{F,M} : \mathbb{Z}_{\geq 0} \to \mathbb{R}$ defined by

$$\mu_{F,M}(n) = \frac{1}{G_n} \sum_{\partial(g)=n} (\widetilde{F}(g) - 1)^M \tag{26}$$

for all $n \geq 0$.

Remark 22. Let F be a Warlimont function on an additive arithmetical semigroup $(G, +, \partial)$. The average value of F on $\partial^{-1}(n)$ is given by

$$\frac{F_n}{G_n} = F_1^+ (1 + \mu_{F,1}(n)).$$

The higher normalized moments can be thought of as capturing the deviation of F from F_1^+. For instance, if $\mu_{F,1}(n) = o(1)$, then

$$\frac{1}{G_n} \sum_{\partial(g)=n} (F(g) - F_1^+)^2 = (F_1^+)^2 \mu_{F,2}(n)$$

can be thought of as an asymptotic measure of the variance of F on $\partial^{-1}(n)$.

Corollary 23. *Let F be a Warlimont function on an additive arithmetical semigroup $(G, +, \partial)$ that satisfies axiom W_1. Then*

$$\lim_{n\to\infty} \frac{F_n}{G_n} = F_1^+,$$

$$\lim_{n\to\infty} \frac{1}{G_n} \sum_{\partial(g)=n} (F(g) - F_1^+)^2 = 0.$$

Proof. Combining Remark 22 and Theorem 20 (with $R = M = 1$), we obtain

$$\frac{F_n}{G_n} = F_1^+ (1 + \mu_{F,1}(n)) = F_1^+ + o(1).$$

Similarly,

$$\frac{1}{G_n} \sum_{\partial(g)=n} (F(g) - F_1^+)^2 = (F_1^+)^2 \left(\xi_1 \frac{G_{n-1}}{G_n} + O\left(\frac{G_{n-1}}{G_n} \right) \right) = o(1). \qquad \square$$

Remark 24. A slightly stronger (see Remark 11) version of Corollary 23 is proved in [Knopfmacher 1976] for particular choices of F. A sharper result is given in [Warlimont 1993] where it is shown that the assumption $G_{n-1} = O(G_n)$ (which is part of axiom $\mathcal{W}_1$) is unnecessary.

Theorem 25. *Let F be a Warlimont function on a Wright semigroup $(G, +, \partial)$ with α, a, b as in Definition 9 and let $q = e^{\alpha(a+1)}$:*

(i) *For every positive integer M there exists a sequence $\{\lambda_s\}$ of functions $\lambda_s : \mathbb{Z}_{\geq 0} \to \mathbb{R}$ such that $\log(\lambda_s(n)) = O(n^{a-1} + n^b)$ and, for every fixed positive integer R, the asymptotic relation*

$$\mu_{F,M}(n) = \sum_{s=1}^{R-1} \lambda_s(n) q^{-sn^a} + O(\lambda_R(n) q^{-Rn^a}) \tag{27}$$

holds in the limit $n \to \infty$.

(ii) *Assume further that there exist constants $0 \leq d_2 \leq d_1$ and a sequence $\{\psi_s\}$ of polynomials such that $\deg(\psi_s) \leq d_1 s - d_2$ for all $s \geq 1$ and, for every fixed positive integer R, the asymptotic relation*

$$\frac{G_{n-1}}{G_n} = \sum_{s=1}^{R-1} \psi_s(n) q^{-ns} + O(n^{d_1 R - d_2} q^{-Rn}) \tag{28}$$

holds in the limit $n \to \infty$. Then there exists a sequence $\{\tau_s\}$ of polynomials such that $\deg(\tau_s) \leq d_1 s - d_2$ and, for every positive integer R, the asymptotic relation

$$\mu_{F,M}(n) = \sum_{s=1}^{R-1} \tau_s(n) q^{-sn} + O(n^{d_1 R - d_2} q^{-Rn}) \tag{29}$$

holds in the limit $n \to \infty$.

Proof. Let ξ_s be defined by (25) for all $s \geq 1$. By Proposition 12 and Theorem 20, we obtain

$$\mu_{F,M}(n) = \sum_{s=1}^{R-1} \xi_s \frac{G_{n-s}}{G_n} + O\left(\frac{G_{n-R}}{G_n}\right) \tag{30}$$

for every fixed integer $R > 0$. Since

$$\log\left(\frac{G_{n-s}}{G_n}\right) = \alpha((n-s)^{a+1} - n^{a+1}) + O(n^b) = -\alpha(a+1)sn^a + O(n^{a-1} + n^b),$$

in order to prove (i) it suffices to choose λ_s such that

$$\lambda_s(n) = \xi_s q^{sn^a} \frac{G_{n-s}}{G_n}$$

for all $n \geq s \geq 1$. Using (28) repeatedly and induction on t, for every fixed positive integer R, we obtain

$$\frac{G_{n-t}}{G_n} = \frac{G_{n-1}}{G_n} \cdots \frac{G_{n-t}}{G_{n-t+1}} = \sum_{s=t}^{R-1} \nu_{s,t}(n) q^{-sn} + O(n^{DR} q^{-Rn}), \tag{31}$$

where

$$\nu_{s,t}(n) = \sum_{i_1+\cdots+i_t=s} \psi_{i_1}(n)\psi_{i_2}(n-1)\cdots\psi_{i_t}(n-t+1)q^{i_2+2i_3+\cdots+(t-1)i_t} \tag{32}$$

is a polynomial in n of degree at most $d_1 s - d_2 t$ for all $1 \le t \le s$. Substituting (31) into (30) yields, for every fixed positive integer R,

$$\begin{aligned}\mu_{F,M}(n) &= \sum_{t=1}^{R-1} \xi_t \sum_{s=t}^{R-1} \nu_{s,t}(n) q^{-ns} + O(n^{DR} q^{-nR}) \\ &= \sum_{s=1}^{R-1} \left(\sum_{t=1}^{s} \xi_t \nu_{s,t}(n) \right) q^{-sn} + O(n^{DR} q^{-nR}),\end{aligned}$$

which proves (ii) upon setting

$$\tau_s(n) = \sum_{t=1}^{s} \xi_t \nu_{s,t}(n) \tag{33}$$

for all s, n. □

Remark 26. Comparison of (28) and (16) shows that the assumptions of (ii) in Theorem 25 require in particular that (16) holds with $a = 1$.

4. Examples

4.1. ***Graphs.*** Let $(G, +)$ be the semigroup of (simple, unlabeled) graphs with semigroup operation $+$ given by disjoint union. If ∂ is the map that to each graph g assigns the cardinality of its set of vertices, then $(G, +, \partial)$ is an additive arithmetical semigroup and $g \in G^+$ if and only if the graph g is connected. As proved in [Wright 1969], there exists a sequence $\{\varphi_s\}$ of polynomials such that φ_s has degree $2s$ for every s and, for every fixed positive integer R, the asymptotic relation

$$G_n = \frac{2^{\binom{n}{2}}}{n!} \left(\sum_{s=0}^{R-1} \varphi_s(n) 2^{-sn} + O(n^{2R} 2^{-Rn}) \right) \tag{34}$$

holds in the limit $n \to \infty$. The polynomials φ_s can be calculated explicitly, the first few being

$$\begin{aligned}\varphi_0(n) &= 1, \\ \varphi_1(n) &= 2n^2 - 2n, \\ \varphi_2(n) &= 8n^4 - \tfrac{128}{3}n^3 + 72n^2 - \tfrac{112}{3}n, \\ \varphi_3(n) &= \tfrac{256}{3}n^6 - \tfrac{3712}{3}n^5 + \tfrac{20672}{3}n^4 - \tfrac{54272}{3}n^3 + 21952n^2 - 9600n.\end{aligned}$$

In particular,

$$\log(G_n) = \log(\sqrt{2})n^2 - n\log(n) + (1 - \log(\sqrt{2}))n + O(n^b)$$

for any $b > 0$ and thus $(G, +, \partial)$ is a Wright semigroup. Moreover, using (34) and expanding the denominator as a geometric series we obtain, for every fixed $R > 0$,

$$\frac{G_{n-1}}{G_n} = 2n2^{-n}\left(\sum_{s=0}^{R-1} 2^s\varphi_s(n-1)2^{-sn}\right)\sum_{r=0}^{R-1}\left(-\sum_{s=1}^{R-1}\varphi_s(n)2^{-sn}\right)^r + O(n^{2R+1}2^{-(R+1)s})$$
$$= \sum_{s=1}^{R-1}\psi_s(n)2^{-sn} + O(n^{2R-1}2^{-Rn}),$$

where the ψ_s are polynomials of degree $\deg(\psi_s) = 2s-1$ which can be explicitly calculated in terms of the polynomials φ_s in (34). For instance

$$\psi_1(n) = n,$$
$$\psi_2(n) = 4n^3 - 20n^2 + 16n,$$
$$\psi_3(n) = 40n^5 - 464n^4 + 1768n^3 - 2624n^2 + 1280n,$$
$$\psi_4(n) = \tfrac{3248}{3}n^7 - 24176n^6 + \tfrac{630608}{3}n^5 - 908496n^4 + \tfrac{6137792}{3}n^3 - 2250240n^2 + 925696n.$$

Substitution into (32) yields $\nu_{s,1}(n) = \psi_s(n)$ for all s and

$$\nu_{2,2}(n) = 8n^2 - 8n,$$
$$\nu_{3,2}(n) = 48n^4 - 352n^3 + 688n^2 - 384n,$$
$$\nu_{3,3}(n) = 64n^3 - 192n^2 + 128n,$$
$$\nu_{4,2}(n) = 864n^6 - 13472n^5 + 77216n^4 - 203488n^3 + 245376n^2 - 106496n,$$
$$\nu_{4,3}(n) = 896n^5 - 9728n^4 + 35200n^3 - 50944n^2 + 24576n,$$
$$\nu_{4,4}(n) = 1024n^4 - 6144n^3 + 11264n^2 - 6144n.$$

Inspection of graphs with up to four vertices shows that $G_1 = 1$, $G_2 = 2$, $G_3 = 4$ and $G_4 = 11$. Substitution into (19) yields $\beta_1 = \beta_2 = \beta_3 = -1$ and $\beta_4 = -4$.

Example 27. Consider the Warlimont function d_2 from Example 15. When specialized to the semigroup of graphs, d_2 counts the number of ways of writing a given graph as the disjoint union of two graphs. The order is taken into account, so that if g_1 is not isomorphic to g_2, then $g = g_1 + g_2$ and $g = g_2 + g_1$ count as two distinct decompositions. Moreover, decompositions in which one of the components is the empty graph are allowed. Combining Remark 22 and Theorem 25 we obtain (4). In particular, setting $M = 1$ yields a full asymptotic expansion for the average of d_2 of the form

$$\frac{1}{G_n}\sum_{\partial(g)=n} d_2(g) = 2 + 2\sum_{s=1}^{R-1}\tau_s(n)2^{-sn} + O(n^{2R-1}2^{-Rn}),$$

valid for every fixed positive integer R, where the $\tau_s(n)$ are polynomials of degree $2s-1$. For instance, direct inspection of graphs with up to four vertices yields

$(d_2)_1 = 2,\ (d_2)_2 = 5,\ (d_2)_3 = 12$ and $(d_2)_4 = 34$. Substituting into (25) and then into (33) we obtain

$$\tau_1(n) = 2n,$$
$$\tau_2(n) = 4n^3 - 4n^2,$$
$$\tau_3(n) = 40n^5 - 368n^4 + 1320n^3 - 2016n^2 + 1024n,$$
$$\tau_4(n) = \tfrac{3248}{3}n^7 - 22448n^6 + \tfrac{560528}{3}n^5 - 781712n^4 + \tfrac{5136512}{3}n^3 - 1839360n^2 + 743424n.$$

4.2. ***Graphs with an even number of edges.*** Let $(G, +)$ be the semigroup of (simple, unlabeled) graphs with an even number of edges and semigroup operation $+$ given by disjoint union. If ∂ is the map that to each graph g assigns the cardinality of its set of vertices, then $(G, +, \partial)$ is an additive arithmetical semigroup. G^+ consists of graphs g with an even number of edges that cannot be written as the disjoint union of two nonempty graphs with an even number of edges. While G is a subsemigroup of the semigroup of all unlabeled graphs, not all graphs in G^+ are connected. For instance, while $2K_1$ is not connected, it is nevertheless prime in the semigroup of graphs with even edges. As pointed out in [Aldi 2019], for every fixed positive integer R, the asymptotic relation

$$G_n = \frac{2^{\binom{n}{2}}}{2n!}\left(\sum_{s=0}^{R-1} \varphi_s(n) 2^{-sn} + O(n^{2R} 2^{-Rn})\right)$$

holds in the $n \to \infty$ limit, where the polynomials $\varphi_s(n)$ coincide with those of Section 4.1. In particular, $(G, +, \partial)$ is a Wright semigroup and, for every fixed positive integer R,

$$\frac{G_{n-1}}{G_n} = \sum_{s=1}^{R-1} \psi_s(n) 2^{-sn} + O(n^{2R-1} 2^{-Rn}),$$

where the polynomials $\psi_s(n)$ coincide with those calculated in Section 4.1. Inspection of graphs with up to four vertices shows that $G_1 = G_2 = 1,\ G_3 = 2$ and $G_4 = 6$. Substitution into (19) yields $\beta_1 = -1,\ \beta_2 = 0,\ \beta_3 = -1$ and $\beta_4 = -3$.

Example 28. Consider the Warlimont function d_* from Example 16. Combining Remark 22 and Theorem 25 we obtain a full asymptotic expansion for the second moment of d_* about 2:

$$\frac{1}{G_n} \sum_{\partial(g)=n} (d_*(g) - 2)^2 = 4 \sum_{s=1}^{R-1} \tau_s(n) 2^{-sn} + O(n^{2R-1} 2^{-Rn})$$

for every fixed positive integer R, where the $\tau_s(n)$ are polynomials of degree $2s - 1$. To calculate these explicitly for small values of s, we first observe by direct calculation that $(d_*)_1 = 2,\ (d_*)_2 = 2,\ (d_*)_3 = 4,\ (d_*)_4 = 14$ as well as $(d_*^2)_1 = 4$, $(d_*^2)_2 = 4,\ (d_*^2)_3 = 8,\ (d_*^2)_4 = 36$. Substitution into (25) (upon setting $M = 2$) and

then into (33) yields

$$\tau_1(n) = 2n,$$
$$\tau_2(n) = 4n^3 - 20n^2 + 16n,$$
$$\tau_3(n) = 40n^5 - 464n^4 + 1832n^3 - 2816n^2 + 1408n,$$
$$\tau_4(n) = \tfrac{3248}{3}n^7 - 24176n^6 + \tfrac{633296}{3}n^5 - 906960n^4 + \tfrac{6040640}{3}n^3 - 2177280n^2 + 882688n.$$

4.3. ***Polynomials over a finite field.*** Consider the field $\mathbb{F}_q$ with q elements and let G be the set of nonzero polynomials in $\mathbb{F}_q[x_1, \dots, x_k]$ modulo the equivalence relation such that $f \sim g$ if and only if $f = \lambda g$ for some $\lambda \in \mathbb{F}_q$. G has a natural structure of additive semigroup with semigroup operation $+$ given by multiplication of polynomials. If ∂ is the semigroup homomorphism that to each polynomial $f \in G$ assigns its total degree, then $(G, +, \partial)$ is an additive arithmetical semigroup and G^+ is the set of equivalent classes of irreducible polynomials in $\mathbb{F}_q[x_1, \dots, x_k]$. Since

$$G_n = \frac{q^{\binom{n+k}{k}} - q^{\binom{n-1+k}{k}}}{q-1} \tag{35}$$

for every n,

$$\log(G_n) = \log(q)\frac{n^k}{k!} + O(n^{k-1})$$

for every $k \geq 2$. On the other hand if $k = 1$, then $\log(G_n) = \log(q)n$ for every n. Hence $(G, \cdot, \partial)$ is a Wright semigroup if and only if $k \geq 2$. If $k = 2$ then for every fixed positive integer R

$$\frac{G_{n-1}}{G_n} = q^{-n-1}\frac{1-q^{-n}}{1-q^{-n-1}} = \sum_{s=1}^{R-1} \psi_s(n) q^{-sn} + O(q^{-Rn}),$$

where $\psi_1(n) = q^{-1}$ and $\psi_s(n) = q^{-s}(1-q)$ for all $s \geq 2$. By Theorem 25, each $\mu_{F,M}$ admits an asymptotic expansion as a power series in q^{-n} with *constant* coefficients. For instance, substitution into (32) yields

$$\nu_{2,1}(n) = q^{-2} - q^{-1}, \quad \nu_{2,2}(n) = q^{-1},$$
$$\nu_{3,1}(n) = q^{-3} - q^{-2}, \quad \nu_{3,2}(n) = q^{-2} - 1, \quad \nu_{3,3}(n) = 1.$$

Example 29. We further specialize to the case where G is the semigroup of nonzero polynomials in two variables over the field with two elements. By Theorem 25, there exist constants τ_s such that for every fixed positive integer R the average of the Warlimont function B (as defined in Example 17) on polynomials of degree n is

$$\frac{B_n}{G_n} = 1 + \sum_{s=1}^{R-1} \tau_s 2^{-sn} + O(2^{-Rn}). \tag{36}$$

Since $B_1 = 6$, $B_2 = 62$ and $B_3 = 1002$, substituting (35) into (19) and then into (25) shows that in particular $\tau_1 = 0$, $\tau_2 = 3$ and $\tau_3 = \frac{3}{2}$.

Example 30. If $k > 2$, then by Remark 26 we are in the second part of Theorem 25. Nevertheless, the asymptotic behavior of Warlimont functions can be described using (27) as follows. Consider for instance the Warlimont function B of Example 17 on the semigroup of polynomials in three variables with coefficients in $\mathbb{F}_q$. Since

$$G_1 = -\beta_1 = (B^m)_1$$

for all m, substitution in (25) yields $\xi_1 = 0$ and thus

$$\frac{1}{G_n} \sum_{\partial(g)=n} (B(g) - 1)^M = O\left(\frac{G_{n-2}}{G_n}\right) = O(q^{-n^2-2n})$$

for all M.

Acknowledgments

Part of this work was carried out during the summer of 2016 at VCU and supported by a UROP Summer Research Fellowship.

References

[Aldi 2019] M. Aldi, "Arithmetical semirings", *Discrete Math.* **342**:7 (2019), 2035–2047. MR Zbl

[Burris 2001] S. N. Burris, *Number theoretic density and logical limit laws*, Math. Surveys and Monographs **86**, Amer. Math. Soc., Providence, RI, 2001. MR Zbl

[Flajolet and Sedgewick 2009] P. Flajolet and R. Sedgewick, *Analytic combinatorics*, Cambridge Univ. Press, 2009. MR Zbl

[Knopfmacher 1975] J. Knopfmacher, *Abstract analytic number theory*, North-Holland Math. Library **12**, North-Holland, Amsterdam, 1975. MR Zbl

[Knopfmacher 1976] J. Knopfmacher, "Arithmetical properties of finite graphs and polynomials", *J. Combinatorial Theory Ser. B* **20**:3 (1976), 205–215. MR Zbl

[Warlimont 1993] R. Warlimont, "A relationship between two sequences and arithmetical semigroups", *Math. Nachr.* **164** (1993), 201–217. MR Zbl

[Warlimont 2001] R. Warlimont, "About the radius of convergence of the zeta function of an additive arithmetical semigroup", *Quaest. Math.* **24**:3 (2001), 355–362. MR Zbl

[Wright 1967] E. M. Wright, "A relationship between two sequences", *Proc. London Math. Soc.* (3) **17** (1967), 296–304. MR Zbl

[Wright 1969] E. M. Wright, "The number of graphs on many unlabelled nodes", *Math. Ann.* **183** (1969), 250–253. MR Zbl

[Wright 1970] E. M. Wright, "Asymptotic relations between enumerative functions in graph theory", *Proc. London Math. Soc.* (3) **20** (1970), 558–572. MR Zbl

Received: 2017-07-08 Revised: 2019-01-08 Accepted: 2019-04-02

maldi2@vcu.edu *Department of Mathematics and Applied Mathematics, Virginia Commonwealth University, Richmond, VA, United States*

tanh4@vcu.edu *Virginia Commonwealth University, Richmond, VA, United States*

INVOLVE 12:7 (2019)
dx.doi.org/10.2140/involve.2019.12.1099

A systematic development of Jeans' criterion with rotation for gravitational instabilities

Kohl Gill, David J. Wollkind and Bonni J. Dichone

(Communicated by Martin J. Bohner)

An inviscid fluid model of a self-gravitating infinite expanse of a uniformly rotating adiabatic gas cloud consisting of the continuity, Euler's, and Poisson's equations for that situation is considered. There exists a static homogeneous density solution to this model relating that equilibrium density to the uniform rotation. A systematic linear stability analysis of this exact solution then yields a gravitational instability criterion equivalent to that developed by Sir James Jeans in the absence of rotation instead of the slightly more complicated stability behavior deduced by Subrahmanyan Chandrasekhar for this model with rotation, both of which suffered from the same deficiency in that neither of them actually examined whether their perturbation analysis was of an exact solution. For the former case, it was not and, for the latter, the equilibrium density and uniform rotation were erroneously assumed to be independent instead of related to each other. Then this gravitational instability criterion is employed in the form of Jeans' length to show that there is very good agreement between this theoretical prediction and the actual mean distance of separation of stars formed in the outer arms of the spiral galaxy Andromeda M31. Further, the uniform rotation determined from the exact solution relation to equilibrium density and the corresponding rotational velocity for a reference radial distance are consistent with the spectroscopic measurements of Andromeda and the observational data of the spiral Milky Way galaxy.

1. Introduction and formulation of the problem

Consider the governing equations for a self-gravitational adiabatic inviscid fluid of infinite extent undergoing uniform rotation [Chandrasekhar 1961]:

$$\text{continuity equation:} \quad \frac{D\rho}{Dt}+\rho\nabla\cdot\boldsymbol{v}=0, \tag{1-1a}$$

MSC2010: 35B36, 35Q85, 76E07, 76E99.

Keywords: Andromeda and Milky Way star formation, Jeans' self-gravitational instabilities, rotating adiabatic inviscid gas dynamics, astrophysics.

$$\text{Euler's equation:} \quad \frac{D\boldsymbol{v}}{Dt}+2\boldsymbol{\Omega}\times\boldsymbol{v}+\boldsymbol{\Omega}\times(\boldsymbol{\Omega}\times\boldsymbol{r}) = -\frac{1}{\rho}\mathscr{P}'(\rho)\nabla\rho+\boldsymbol{g}, \tag{1-1b}$$

$$\text{Poisson's equation:} \quad \nabla\cdot\boldsymbol{g} = -4\pi G_0\rho. \tag{1-1c}$$

Here $t \equiv$ time, $\boldsymbol{r} = (x, y, z) \equiv$ position vector, $\boldsymbol{\Omega} = (0, 0, \Omega_0) \equiv$ uniform rotation vector, $\rho \equiv$ density (mass/[unit volume]), $\boldsymbol{v} = (u, v, w) \equiv$ velocity vector with respect to the rotating frame, $\nabla = (\partial/\partial x, \partial/\partial y, \partial/\partial z) \equiv$ gradient operator, $D/Dt = \partial/\partial t+\boldsymbol{v}\cdot\nabla \equiv$ material derivative, $\mathscr{P}(\rho) = p_0(\rho/\rho_0)^{\gamma_0} \equiv$ adiabatic pressure, $\boldsymbol{g} = -\nabla\varphi \equiv$ gravitational acceleration vector with $\varphi \equiv$ self-gravitating potential, and $G_0 \equiv$ universal gravitational constant. The continuity and Euler's equations follow from the conservation of mass and momentum for an inviscid fluid [Lin and Segel 1974] with the addition of the extra second and third terms on the left-hand side of (1-1b), which represent the Coriolis effect and centrifugal force, respectively, due to the rotation [Greenspan 1968]. Poisson's equation follows from the divergence theorem and Newton's law of universal gravitation [Binney and Tremaine 1987; Lin and Segel 1974]. Since

$$\begin{gathered}\boldsymbol{\Omega}\times(\boldsymbol{\Omega}\times\boldsymbol{r}) = (\boldsymbol{\Omega}\cdot\boldsymbol{r})\boldsymbol{\Omega}-(\boldsymbol{\Omega}\cdot\boldsymbol{\Omega})\boldsymbol{r} = -\Omega_0^2(x, y, 0),\\ \boldsymbol{\Omega}\times\boldsymbol{v} = \Omega_0(-v, u, 0), \quad \nabla\cdot\boldsymbol{g} = -\nabla^2\varphi\end{gathered} \tag{1-1d}$$

[Segel 1977], the Euler's and Poisson's equations become

$$\frac{D\boldsymbol{v}}{Dt}+2\Omega_0(-v,u,0)-\Omega_0^2(x,y,0) = -\frac{1}{\rho}\mathscr{P}'(\rho)\nabla\rho-\nabla\varphi, \tag{1-1e}$$

$$\nabla^2\varphi = 4\pi G_0\rho, \quad \text{where } \nabla^2 \equiv \nabla\cdot\nabla. \tag{1-1f}$$

Sir James Jeans [1902; 1928] proposed that a gravitational instability mechanism occurring in the spiral arms of protogalactic nebulae could result in the formation of chains of condensations, which eventually developed into those stars visible in the outer regions of fully evolved galaxies. He suggested that a nonrotating self-gravitating unbounded interstellar cloud of adiabatic gas, which is initially uniform in density and quiescent, should undergo an instability mechanism of this sort when acted on by random infinitesimal perturbations. Jeans [1902; 1928] deduced a criterion for which such an interstellar cloud would exhibit a gravitational instability by performing a linear stability analysis on what he assumed to be an exact solution to his governing inviscid gas dynamical model system that was equivalent to equations (1-1) in the absence of rotation, arriving at the following secular equation satisfied by σ and λ, the growth rate and wavelength, respectively, of his small density fluctuations:

$$\sigma^2 = 4\pi\left(G_0\rho_0 - \pi\frac{c_0^2}{\lambda^2}\right), \tag{1-2a}$$

where c_0 is the speed of sound in an adiabatic medium of uniform density ρ_0. This relation differed from that for the propagation of sound in a homogeneous medium

only due to the presence of the gravity term in (1-2a). Then, from (1-2a), Jeans concluded that there would be instability corresponding to $\sigma^2 > 0$ provided

$$\lambda > \lambda_J = c_0\sqrt{\frac{\pi}{G_0\rho_0}} \equiv \text{Jeans' length}, \tag{1-2b}$$

which is known as Jeans' criterion for gravitational instabilities.

The only problem with this derivation is that Jeans represented his exact static solution to those governing equations symbolically as $\boldsymbol{v} \equiv \mathbf{0} = (0, 0, 0)$, $\rho = \rho_0$, $\varphi = \varphi_0$. Since this analysis was for a nonrotating system with $\Omega_0 = 0$, when he assumed in addition that ρ_0 was uniform to make his perturbation equations constant coefficient this implicitly required $\boldsymbol{\nabla}\varphi_0 = \mathbf{0}$, which implied $\nabla^2\varphi_0 = 0 = 4\pi G_0\rho_0$ or $\rho_0 = 0$ and hence is termed Jeans' swindle by Binney and Tremaine [1987]. Kiessling [2003] refutes their claim that Jeans' derivation represents a swindle because it can be justified by taking the proper limit of the appropriate cosmological model.

Since spectroscopic evidence (reviewed by Rubin and Ford [1970]) ultimately showed these nebulae to be rotating, Subrahmanyan Chandrasekhar [1961] considered the effect of adding rotation to Jeans' governing system of perturbation equations and repeated that analysis, demonstrating in the process that its stability behavior was slightly more complicated in that it involved an extra instability condition as well as Jeans' criterion. Chandrasekhar's perturbation analysis suffered from the same deficiency as Jeans' in that he did not develop a parameter relationship for his implicit exact solution and thus treated ρ_0 and Ω_0 as independent. We shall demonstrate that the proper relationship between these parameters eliminates this extra condition and only yields Jeans' instability criterion. Many subsequent linear stability analyses of similar problems influenced by the methodology of these works have treated their associated perturbation systems independently of the actual exact solution of the governing equations and thus replicate this deficiency including recent studies and reviews of gravitational instabilities [Stahler and Palla 2004]. Hence, we believe there is some merit in performing a systematic linear stability analysis of the relevant exact solution for Chandrasekhar's problem and toward that end present an investigation of this sort in the next section.

2. The exact static homogeneous density solution and its linear stability

There exists an exact static homogeneous density solution of our basic equations of the form

$$\boldsymbol{v} \equiv \mathbf{0} = (0, 0, 0), \quad \rho \equiv \rho_0, \quad \varphi = \varphi_0, \tag{2-1a}$$

where φ_0 satisfies

$$\boldsymbol{\nabla}\varphi_0 = \Omega_0^2(x, y, 0), \quad \nabla^2\varphi_0 = 4\pi G_0\rho_0 \tag{2-1b}$$

or

$$\varphi_0(x, y) = \tfrac{1}{2}\Omega_0^2(x^2 + y^2), \quad \text{with } \Omega_0^2 = 2\pi G_0 \rho_0 > 0. \tag{2-1c}$$

Now seeking a linear perturbation solution of these basic equations of the form

$$\begin{aligned} \boldsymbol{v} &= \varepsilon \boldsymbol{v}_1 + \boldsymbol{O}(\varepsilon^2), \quad \text{where } \boldsymbol{v}_1 = (u_1, v_1, w_1), \\ \rho &= \rho_0[1 + \varepsilon s + O(\varepsilon^2)], \quad \varphi = \varphi_0 + \varepsilon\varphi_1 + O(\varepsilon^2), \end{aligned} \tag{2-2}$$

with $|\varepsilon| \ll 1$, substituting (2-2) into those equations, neglecting terms of $O(\varepsilon^2)$, and canceling the resulting ε common factor; we deduce that the perturbation quantities to this exact solution satisfy

$$\frac{\partial s}{\partial t} + \frac{\partial u_1}{\partial x} + \frac{\partial v_1}{\partial y} + \frac{\partial w_1}{\partial z} = 0, \tag{2-3a}$$

$$\frac{\partial u_1}{\partial t} - 2\Omega_0 v_1 + c_0^2 \frac{\partial s}{\partial x} + \frac{\partial \varphi_1}{\partial x} = 0, \quad \text{where } c_0^2 = \mathscr{P}'(\rho_0) = \gamma_0 \frac{p_0}{\rho_0} > 0, \tag{2-3b}$$

$$\frac{\partial v_1}{\partial t} + 2\Omega_0 u_1 + c_0^2 \frac{\partial s}{\partial y} + \frac{\partial \varphi_1}{\partial y} = 0, \tag{2-3c}$$

$$\frac{\partial w_1}{\partial t} + c_0^2 \frac{\partial s}{\partial z} + \frac{\partial \varphi_1}{\partial z} = 0, \tag{2-3d}$$

$$2\Omega_0^2 s - \nabla^2 \varphi_1 = 0. \tag{2-3e}$$

Then assuming a normal mode solution for these perturbation quantities of the form

$$[u_1, v_1, w_1, s, \varphi_1](x, y, z, t) = [A, B, C, E, F] e^{i(k_1 x + k_2 y + k_3 z) + \sigma t}, \tag{2-4}$$

where $|A|^2 + |B|^2 + |C|^2 + |E|^2 + |F|^2 \neq 0$, $i = \sqrt{-1}$, and $k_{1,2,3} \in \mathbb{R}$ satisfy the implicit far-field boundedness property for those quantities, and substituting (2-4) into (2-3), we obtain the following equations for $[A, B, C, E, F]$ upon cancellation of the exponential common factor:

$$ik_1 A + ik_2 B + ik_3 C + \sigma E = 0, \tag{2-5a}$$

$$\sigma A - 2\Omega_0 B + ic_0^2 k_1 E + ik_1 F = 0, \tag{2-5b}$$

$$2\Omega_0 A + \sigma B + ic_0^2 k_2 E + ik_2 F = 0, \tag{2-5c}$$

$$\sigma C + ic_0^2 k_3 E + ik_3 F = 0, \tag{2-5d}$$

$$2\Omega_0^2 E + k^2 F = 0, \quad \text{where } k^2 = k_1^2 + k_2^2 + k_3^2. \tag{2-5e}$$

Setting the determinant of the 5×5 coefficient matrix for the linear homogeneous system (2-5) of constants equal to zero to satisfy their nontriviality property, we obtain

$$k^2[\sigma^4 + (c_0^2 k^2 + 2\Omega_0^2)\sigma^2] + 4\Omega_0^2(c_0^2 k^2 - 2\Omega_0^2)k_3^2 = 0. \tag{2-6}$$

Defining the wavenumber vector $\boldsymbol{k} = (k_1, k_2, k_3)$, its dot product with $\boldsymbol{\Omega}$ satisfies

$$\boldsymbol{k} \cdot \boldsymbol{\Omega} = k_3 \Omega_0 = |\boldsymbol{k}|\,|\boldsymbol{\Omega}| \cos(\theta) = k \Omega_0 \cos(\theta), \tag{2-7a}$$

θ being the azimuthal angle between $\boldsymbol{k}$ and $\boldsymbol{\Omega}$, which implies

$$k_3 = k \cos(\theta). \tag{2-7b}$$

Then, substitution of (2-7b) into (2-6) and cancellation of k^2 yields the secular equation

$$\sigma^4 + (c_0^2 k^2 + 2\Omega_0^2)\sigma^2 + 4\Omega_0^2 (c_0^2 k^2 - 2\Omega_0^2) \cos^2(\theta) = 0. \tag{2-8}$$

Since this secular equation is a quadratic in σ^2, we first demonstrate that $\sigma^2 \in \mathbb{R}$ by showing that its discriminant $\mathcal{D}$ satisfies

$$\mathcal{D} = (c_0^2 k^2 + 2\Omega_0^2)^2 - 16\Omega_0^2 (c_0^2 k^2 - 2\Omega_0^2) \cos^2(\theta) \geq 0. \tag{2-9a}$$

Consider the two cases of $c_0^2 k^2 - 2\Omega_0^2 \leq 0$ and $c_0^2 k^2 - 2\Omega_0^2 > 0$ separately. For the former case it is obvious, while for the latter one it can be deduced by noting that

$$\mathcal{D} \geq (c_0^2 k^2 + 2\Omega_0^2)^2 - 16\Omega_0^2 (c_0^2 k^2 - 2\Omega_0^2) = (c_0^2 k^2 - 6\Omega_0^2)^2. \tag{2-9b}$$

For $\theta = \frac{\pi}{2}$, we can conclude from (2-8) that

$$\sigma^2 = 0 \quad \text{or} \quad \sigma^2 = -(c_0^2 k^2 + 2\Omega_0^2) < 0, \tag{2-10a}$$

while for $\theta \neq \frac{\pi}{2}$, the stability criteria governing such quadratics, namely,

$$\text{given } \omega^2 + a\omega + b = 0 \text{ with } \mathcal{D} = a^2 - 4b \geq 0, \quad \omega < 0 \iff a, b > 0 \tag{2-10b}$$

[Uspensky 1948], implies that

$$\sigma^2 < 0 \iff c_0^2 k^2 - 2\Omega_0^2 > 0. \tag{2-10c}$$

Making an interpretation of these results, we can deduce from (2-10) and (2-1c) that there will only be $\sigma^2 > 0$ and hence unstable behavior provided

$$c_0^2 k^2 - 4\pi G_0 \rho_0 < 0, \tag{2-11a}$$

which is equivalent to Jeans' gravitational instability criterion (1-2b)

$$\lambda > \lambda_J = c_0 \sqrt{\frac{\pi}{G_0 \rho_0}} \equiv \text{Jeans' length} \tag{2-11b}$$

since

$$\lambda = \frac{2\pi}{k}. \tag{2-11c}$$

Figure 1. A Galaxy Evolution Explorer image of the Andromeda galaxy M31, courtesy NASA/JPL-Caltech.

3. Comparisons

Let us return to Jeans' analysis. In writing (1-2), one must implicitly assume that $\rho_0 > 0$, which seems plausible in that $\rho_0 = 0$ corresponds to a completely empty space [Scheffler and Elsässer 1988]. When gravity is taken into account in the absence of rotation, however, such an assumption is not strictly compatible with the equations of hydrostatic equilibrium, as we have seen. Thus, Jeans' uniform density solution, as mentioned above, was not exact. The problem under examination demonstrates that adding rotation to the system as Chandrasekhar did and again performing a standard linear stability analysis of its exact static solution yields Jeans' instability criterion but in a systematic manner and such a model also has the added advantage of being more astrophysically realistic. Jeans got the right answer for the wrong reason, as was shown in [Kiessling 2003] by taking the proper limit of the appropriate cosmological model to fix that analysis. In his review of hydrodynamic stability theory, the renowned comprehensive applied mathematical modeler Lee Segel [1966] stated that "Anyone can get the right answer for the right reason. It takes a genius or a physicist to get the right answer for the wrong reason." In this context, Sir James Jeans was both.

The formula for λ_J in (2-11b) is of fundamental importance in astrophysics and cosmology where many significant deductions concerning the formation of galaxies and stars have been based upon it. In particular, Jeans' interpretation of the criterion, now bearing his name, was that a gas cloud of characteristic dimension much greater than λ_J would tend to form condensations with mean distance of separation comparable to λ_J that then developed into those protostars observable in the outer arms of spiral galaxies such as Andromeda M31 (see Figure 1). Sekimura

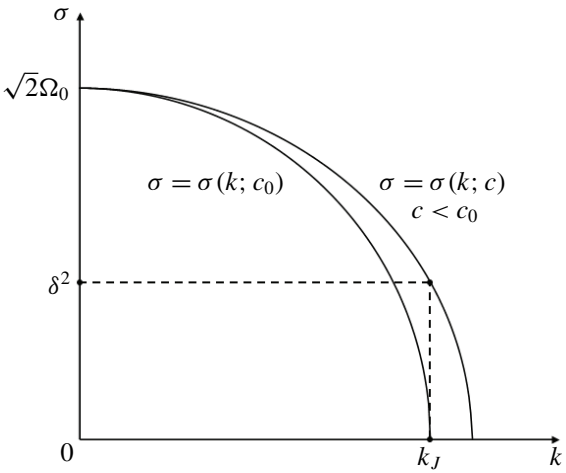

Figure 2. Schematic plots in the k-σ plane depicting the methodology employed by [Sekimura et al. 1999] applied to the Jeans' secular equation $\sigma = \sigma(k; c) = \sqrt{2\Omega_0^2 - c^2k^2}$. That curve is plotted for both a general speed of sound c and our specific speed $c_0 > c$ in this figure where $k_J = 2\pi/\lambda_J$ is such that $\sigma(k_J; c_0) = 0$. In a weakly nonlinear stability analysis one takes the disturbance wavenumber $k \equiv k_J$ and its growth rate to be equal to $\sigma_J(c) = \sigma(k_J; c) = \delta^2 > 0$ where c is close enough to c_0 so that δ is a small parameter. Then in the $\lim_{c\to c_0} \sigma_J(c) = 0$ which is a requirement for the application of weakly nonlinear stability theory and any re-equilibrated pattern will exhibit a wavelength of λ_J. Here $c^2 = \gamma(p_0/\rho_0)$ with $\gamma < \gamma_0$ and hence the operation $\lim_{c\to c_0}$ is equivalent to $\lim_{\gamma\to\gamma_0}$.

et al. [1999] have demonstrated that, for a secular equation similar in form to (1-2a), λ_J actually corresponds to the so-called critical wavelength λ_c of linear stability theory associated with $\sigma = 0$ (see Figure 2), while nonlinear stability analyses of physical phenomena involving related secular equations have shown that the observed wavelengths are determined to a close approximation by that λ_c rather than by the dominant wavelength λ_d at which σ achieves its maximum value from linear theory (see, e.g., [Tian and Wollkind 2003]). Hence, Jeans' interpretation, although unusual for linear stability theory (where it is often presumed that such a disturbance associated with the largest growth rate predominates), both anticipated and is consistent with these nonlinear results, since, by the time perturbations have grown enough for the effect of the maximum growth rate to be observed, the neglected nonlinearities may have rendered that linear analysis inaccurate [Segel and Stoeckly 1972]. In this context, note that for a typical value of $\theta \neq \frac{\pi}{2}$, namely $\theta = 0$, we can factor our secular equation (2-8) to obtain the roots $\sigma^2 = -4\Omega_0^2$ and

$\sigma^2 = 2\Omega_0^2 - c_0^2 k^2$. Observe that this last condition, which yields our instability, is equivalent to the Jeans' secular equation (1-2a).

Thus using the formula for Jeans' length λ_J with the parameters c_0 and ρ_0 assigned the values

$$c_0 = \tfrac{2}{3} \times 10^4 \frac{\text{cm}}{\text{sec}} \quad \text{and} \quad \rho_0 = 10^{-22} \frac{\text{gm}}{\text{cm}^3} \tag{3-1a}$$

employed by [Jeans 1928] for this purpose but when the polytropic index γ_0 is $\frac{4}{3}$ [Bonnor 1957], while taking

$$G_0 = 6.67 \times 10^{-8} \frac{\text{cm}^3}{\text{gmsec}^2} \tag{3-1b}$$

in cgs units yields

$$\lambda_J = 4.58 \times 10^{18}\,\text{cm} = 1.48\,\text{pc}, \tag{3-1c}$$

where 1 pc $\equiv 3.09 \times 10^{18}$ cm, which compares quite favorably with the mean distance between actual adjacent condensations originally formed in the outer arms of Andromeda since, in those parts of M31, the averaged observed distance between protostars in such chains is about 1.4 pc or somewhat more if allowances are made for foreshortening [Jeans 1928].

Given the small size of ρ_0 in (3-1a), Chandrasekhar [1961] was one of those individuals who regarded Jeans' analysis as a close approximation to reality [Schefler and Elsässer 1988]. Although he oriented his axes so that $\boldsymbol{\Omega} = (0, \Omega_y, \Omega_z)$ with $|\boldsymbol{\Omega}| = \Omega_0$ and $\boldsymbol{k} = (0, 0, k)$, using our more general orientation Chandrasekhar, in effect, considered uniform rotation Ω_0 in his perturbation equations through the Coriolis force terms of (2-3b) and (2-3c) in order to make the model more realistic while retaining the coefficient $4\pi\rho_0 G_0$ for s in (2-3e). In so doing, he implicitly assumed that Ω_0 and ρ_0 were independent rather than related parameters. Chandrasekhar plotted σ^2 versus k for $\theta = 0, \frac{\pi}{4}, \frac{\pi}{2}$ and $\Lambda^2 \equiv \Omega_0^2/(\pi G_0 \rho_0) = 0.5, 1.0, 2.0$. Besides Jeans' criterion for $\theta \neq \frac{\pi}{2}$, this yielded an extra extraneous instability criterion for the case of $\theta = \frac{\pi}{2}$, namely,

$$c_0^2 k^2 < 4(\pi G_0 \rho_0 - \Omega_0^2) \quad \text{should } \Omega_0^2 < \pi G_0 \rho_0. \tag{3-2}$$

In point of fact, $\Lambda^2 = 0.5$ is a representative value of that quantity for this instability condition of (3-2), while $\Lambda^2 = 2.0$, his upper bound, actually corresponds to its value as per our formula of (2-1c) relating these parameters, which implies

$$\Omega_0 = \sqrt{2\pi\rho_0 G_0}. \tag{3-3a}$$

Let us examine the plausibility of (3-3a), which violates (3-2) identically. In conjunction with the values for ρ_0 and G_0 of (3-1), (3-3a) yields the uniform

rotation

$$\Omega_0 = 6.47 \times 10^{-15}/\text{sec} \tag{3-3b}$$

and the corresponding rotational velocity

$$V_0 = r_0\Omega_0 = 200\,\frac{\text{km}}{\text{sec}} \tag{3-3c}$$

for the reference radial distance of

$$r_0 = 1\,\text{kpc} = 10^3\,\text{pc} = 3.09 \times 10^{21}\,\text{cm} = 3.09 \times 10^{16}\,\text{km}, \tag{3-3d}$$

both of which are consistent with the spectroscopic measurements of the Andromeda nebula and the observational data of the spiral Milky Way galaxy [Rubin and Ford 1970].

In conclusion our development presents a systematic linear stability analysis of Chandrasekhar's [1961] gravitational instability model in the presence of uniform rotation. We close by noting that Binney and Tremaine [1987] considered this gravitational instability model in a cylindrical rotating system as a problem in Chapter 5 of their book *Galactic Dynamics*. They observed that rotation allowed the Jeans' instability to be analyzed exactly. Since the first part of their problem was to find the condition on Ω_0 so that the homogeneous quiescent gas would be in equilibrium, Binney and Tremaine did not examine the plausibility of this condition. Further, the last part of their problem was to show, upon finding the resulting secular equation from its linear stability analysis, that waves propagating perpendicular to the rotation vector were always stable, while those propagating parallel to it were unstable if and only if the usual Jeans' criterion without rotation was satisfied. Although the latter conclusion for $\theta = 0$ agrees with our predictions, the former does not since, when $\theta = \frac{\pi}{2}$, we predicted $\sigma^2 = 0$, as well as those $\sigma^2 < 0$ which only implies a condition of neutral stability. Our results demonstrate that the best way to test the validity of a model for a natural science phenomenon is to compare its theoretical predictions with observable data of this phenomenon. Sir Arthur Conan Doyle characterized that philosophy probably as well as anyone by a Sherlock Holmes quote from "A scandal in Bohemia" in his 1891 collection entitled *The adventures of Sherlock Holmes*:

> It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts.

References

[Binney and Tremaine 1987] J. Binney and S. Tremaine, *Galactic dynamics*, Princeton Univ. Press, 1987. Zbl

[Bonnor 1957] W. B. Bonnor, "Jeans' formula for gravitational instability", *Monthly Not. Roy. Astr. Soc.* **117** (1957), 104–117. MR Zbl

[Chandrasekhar 1961] S. Chandrasekhar, *Hydrodynamic and hydromagnetic stability*, Int. Series Monogr. Phys. **19**, Clarendon, Oxford, 1961. MR Zbl

[Greenspan 1968] H. P. Greenspan, *The theory of rotating fluids*, Cambridge Monogr. Mech. Appl. Math. **12**, Cambridge Univ. Press, 1968. Zbl

[Jeans 1902] J. H. Jeans, "The stability of a spherical nebula", *Philos. Trans. R. Soc. Lond. Ser. A* **199** (1902), 1–53. Zbl

[Jeans 1928] J. H. Jeans, *Astronomy and cosmogony*, Cambridge Univ. Press, 1928. Zbl

[Kiessling 2003] M. K.-H. Kiessling, "The 'Jeans swindle': a true story — mathematically speaking", *Adv. Appl. Math.* **31**:1 (2003), 132–149. MR Zbl

[Lin and Segel 1974] C. C. Lin and L. A. Segel, *Mathematics applied to deterministic problems in the natural sciences*, Macmillan, New York, 1974. MR Zbl

[Rubin and Ford 1970] V. C. Rubin and W. K. Ford, Jr., "Rotation of the Andromeda Nebula from a spectroscopic survey of emission", *Astrophys. J.* **159** (1970), 379–403.

[Scheffler and Elsässer 1988] H. Scheffler and H. Elsässer, *Physics of the galaxy and interstellar matter*, Springer, 1988.

[Segel 1966] L. A. Segel, "Non-linear hydrodynamic stability theory and its applications to thermal convection and curved flows", pp. 165–197 in *Non-equilibrium thermodynamics, variational techniques and stability* (Chicago, 1965), edited by R. J. Donnelly et al., Univ. Chicago Press, 1966. MR

[Segel 1977] L. A. Segel, *Mathematics applied to continuum mechanics*, Macmillan, New York, 1977. MR Zbl

[Segel and Stoeckly 1972] L. A. Segel and B. Stoeckly, "Instability of a layer of chemotactic cells, attractant and degrading enzyme", *J. Theoret. Biol.* **37**:3 (1972), 561–585.

[Sekimura et al. 1999] T. Sekimura, M. Zhu, J. Cook, P. K. Maini, and J. D. Murray, "Pattern formation of scale cells in Lepidoptera by differential origin-dependent cell adhesion", *Bull. Math. Biol.* **61**:5 (1999), 807–828. Zbl

[Stahler and Palla 2004] S. W. Stahler and F. Palla, *The formation of stars*, Wiley-VCH, Weinheim, Germany, 2004.

[Tian and Wollkind 2003] E. M. Tian and D. J. Wollkind, "A nonlinear stability analysis of pattern formation in thin liquid films", *Interfaces Free Bound.* **5**:1 (2003), 1–25. MR Zbl

[Uspensky 1948] J. V. Uspensky, *Theory of equations*, McGraw-Hill, New York, 1948.

Received: 2018-01-03 Revised: 2019-05-22 Accepted: 2019-05-22

gillkohl@gmail.com *Department of Mathematics, Washington State University, Pullman, WA, United States*

dwollkind@wsu.edu *Department of Mathematics, Washington State University, Pullman, WA, United States*

dichone@gonzaga.edu *Department of Mathematics, Gonzaga University, Spokane, WA, United States*

INVOLVE 12:7 (2019)
dx.doi.org/10.2140/involve.2019.12.1109

The linking-unlinking game

Adam Giambrone and Jake Murphy

(Communicated by Kenneth S. Berenhaut)

Combinatorial two-player games have recently been applied to knot theory. Examples of this include the knotting-unknotting game and the region unknotting game, both of which are played on knot shadows. These are turn-based games played by two players, where each player has a separate goal to achieve in order to win the game. In this paper, we introduce the linking-unlinking game which is played on two-component link shadows. We then present winning strategies for the linking-unlinking game played on all shadows of two-component rational tangle closures and played on a large family of general two-component link shadows.

1. Introduction

Recently, a number of researchers have applied game theory to knot theory in the form of combinatorial games played on knot diagrams. Examples of such games include twist untangle [Ganzell et al. 2014], the knotting-unknotting game [Henrich et al. 2011; Johnson 2011], and the region unknotting game [Brown et al. 2017]. The game twist untangle is played between two people on a nontrivial diagram of the unknot formed by iteratively twisting the unknotted circle. Players take turns using either an R1 move or an R2 move (see Figure 6) to decrease the number of crossings and simplify the diagram. The winner is the player that reduces the diagram to the unknotted circle. For more details, see [Ganzell et al. 2014]. The knotting-unknotting game is played on a *shadow* of a knot (see Definition 2.3). Players take turns *resolving* a crossing (see Definition 2.4) until all crossings are resolved and a knot diagram is formed. One player, the unknotter, wins if the resulting knot diagram is unknotted (represents the trivial knot), while the other player, the knotter, wins if the resulting knot diagram is knotted (represents a nontrivial knot). For more details, see [Henrich et al. 2011; Johnson 2011]. The region unknotting game is similar to the knotting-unknotting game in that the goals of the knotter and unknotter remain the same. The key difference is that play now

MSC2010: 57M25, 91A46.

Keywords: knot, knot diagram, link, link diagram, linking-unlinking game, pseudodiagram, rational link, rational tangle, splittable, two-player game, unsplittable, winning strategy.

consists of resolving the set of crossings that are incident to a face of the knot shadow (or changing the crossing type of a crossing if it has already been resolved). For more details, see [Brown et al. 2017].

In this paper, we will adapt the knotting-unknotting game to be played on two-component link shadows. We call this new game the *linking-unlinking game*. Here, players still take turns resolving crossings. One player, the *unlinker*, wins if the resulting two-component link diagram is *splittable* (is equivalent to a two-component link diagram where the components are separated from each other), while the other player, the *linker*, wins if the resulting two-component link diagram is *unsplittable* (is not splittable).

Our main goal in this paper is to find winning strategies for playing the linking-unlinking game on families of two-component link shadows. The first family of link shadows we explore consists of the shadows of the two-component links that arise as a closure of the rational tangle $(a_1, \ldots, a_n)$. Roughly speaking, a rational tangle is formed by taking two vertical parallel strands and alternating between applying twists to the bottom two endpoints of the strands and applying twists to the right two endpoints of the strands. For an example of the construction of a rational tangle, see Figure 16. For the precise definition of rational tangle, see Definition 2.14. Our first three main results are combined into the following theorem.

Theorem 1.1. *Suppose we have a shadow of a rational two-component link coming from a closure of the rational tangle* $(a_1, \ldots, a_n)$:

(1) *If* $a_{2k+1} = 0$ *for all* k, *then the unlinker wins.*

(2) *If either*

  (a) $a_{2k+1} \neq 0$ *for at least one* k *and all of the* a_i *are even*,

  (b) $n = 2$ *and both* a_1 *and* a_2 *are odd*, *or*

  (c) $n \geq 3$, *both* a_1 *and* a_n *are odd*, *and all other* a_i *are even*,

  then the second player has a winning strategy (*regardless of their role*).

To extend the results above to all shadows of two-component rational tangle closures, we utilize a decomposition of the syllables of the tangle word $(a_1, \ldots, a_n)$ into syllables consisting of self-intersections and strings of syllables consisting of non-self-intersections (see Definition 2.18 and Proposition 2.22). By combining this decomposition with the proof of Theorem 1.1, we are able to prove the following result.

Theorem 1.2. *Suppose we have a shadow of a rational two-component link coming from a closure of the rational tangle* $(a_1, \ldots, a_n)$. *Furthermore, assume* $a_{2k+1} \neq 0$ *for some* k:

(1) *If the tangle word contains an even number of self-intersections, then the second player has a winning strategy* (*regardless of their role*).

(2) *If the tangle word contains an odd number of self-intersections, then the first player has a winning strategy (regardless of their role).*

Finally, we conclude the paper by expanding our focus to general two-component link shadows, using a decomposition of the crossings of a link shadow into self-intersections and non-self-intersections (see Definition 2.11) and using linking number arguments (see Definitions 2.9, 2.10, and 2.12) to prove the following result.

Theorem 1.3. *Suppose we have a shadow of a general two-component link:*

(1) *If the shadow contains zero non-self-intersections, then the unlinker wins.*

(2) *If the shadow contains a nonzero number of non-self-intersections and an even number of self-intersections, then the linker has a winning strategy when playing second.*

(3) *If the shadow contains a nonzero number of non-self-intersections and an odd number of self-intersections, then the linker has a winning strategy when playing first.*

The remainder of this paper is organized as follows. In Section 2, we present background from knot theory, defining terminology and providing results that will be used later. Specifically, we begin by introducing knots and links, various types of knot and link projections, and notions of equivalence for these objects in Section 2A. In Section 2B, we define splittable and unsplittable link diagrams, self-intersections and non-self-intersections for link diagrams, and the linking number of a two-component link diagram. In Section 2C, we define rational tangles and rational link diagrams. In Section 2D, we define self-intersections and non-self-intersections for rational tangles, determine exactly when a rational tangle will close to form a two-component link diagram, and provide a decomposition theorem for rational tangle words.

In Section 3, we define the linking-unlinking game and present winning strategies for playing the game on various two-component link shadows. Specifically, we define the linking-unlinking game and present two key player strategies in Section 3A, we present winning strategies for all shadows of two-component rational tangle closures in Section 3B, and we present winning strategies for large families of general two-component link shadows in Section 3C.

2. Definitions and background

To begin, we introduce some basic ideas from knot theory. Many of the definitions that follow can be found in introductory knot theory textbooks such as [Johnson and Henrich 2017; Adams 1994].

2A. ***Knots, knot projections, and knot equivalence.*** A mathematical knot is much like the everyday knot we tie in our shoelaces. The key difference is that mathematical knots form closed loops.

Figure 1. Some examples of knots.

Definition 2.1. A *knot* is a piecewise-linear simple closed curve in three-dimensional space $\mathbb{R}^3$. Equivalently, a *knot* is a smooth embedding of a circle into $\mathbb{R}^3$.

Returning to our shoelace analogy, if we fused the loose ends of a knotted shoelace together, we would have a model of a mathematical knot. Figure 1 provides some examples of knots. Note that, if we consider the piecewise-linear definition of a knot, we can assume that knots are made up of a very large number of line segments so that the strands of the knot appear smooth. We now shift our attention to links, which consist of knots.

Definition 2.2. A *link* is a collection of one or more knots (that can be, but do not have to be, interlinked). An *n-component link* is a collection of n knots.

Note that a knot is a 1-component link, which means the study of links includes the study of knots. For the majority of this paper, we will focus on two-component links. See Figure 2 for some examples of links with multiple components. To simplify the study of knotted loops in three dimensions, we will carefully project our links to two dimensions.

Definition 2.3. A *link shadow* is a projection of a link onto the plane $\mathbb{R}^2$ so that all crossings are transverse double crossings.

We can think of a link shadow as being formed by carefully shining a flashlight on a link and viewing the shadow it casts. Figure 3 provides both an example and a nonexample of a link shadow.

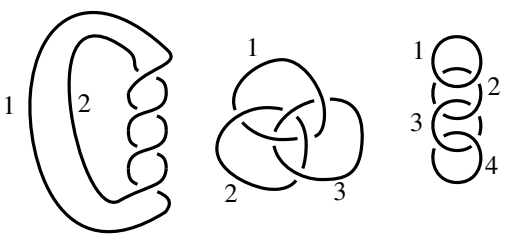

Figure 2. A two-component link (left), a 3-component link (middle), and a 4-component link (right). Note that the components of each link have been numbered.

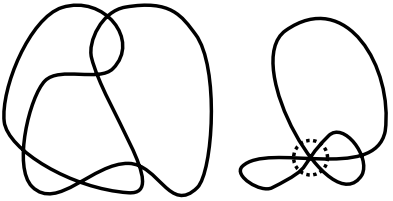

Figure 3. The figure on the left is a valid link shadow. The figure on the right is not, as the circled crossing has three strands meeting at the crossing rather than two.

Notice that link shadows only tell us where crossings occur, not which strand is above the other at each crossing. This means that multiple links can have the same link shadow. To indicate a specific link, we need to include more information.

Definition 2.4. A *resolved crossing* is a crossing in a link projection that has been *resolved* so that the *overstrand* and the *understrand* are distinguishable. This is usually depicted by adding two gaps to the understrand. A crossing that has not been resolved is called an *unresolved crossing*.

To resolve an unresolved crossing, we choose which strand becomes the overstrand. There are two possible resolutions for each crossing, as shown in Figure 4. We now define the results of resolving a subset of crossings of a link shadow.

Definition 2.5. A *link pseudodiagram*, as defined in [Hanaki 2010], is a link projection where an arbitrary number of crossings have been resolved. From this perspective, a link shadow is a link pseudodiagram where no crossings have been resolved and a *link diagram* is a link pseudodiagram where every crossing has been resolved.

For an example of a link pseudodiagram, see Figure 5. Given the family of all links and the family of all link diagrams, we will now discuss notions of equivalence for each of these families.

Definition 2.6. If we can manipulate three-dimensional space $\mathbb{R}^3$ to deform one link L_1 into another link L_2, then L_1 and L_2 are called *equivalent*.

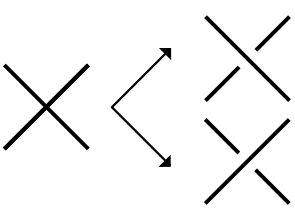

Figure 4. An unresolved crossing of a link projection (left) and the two possible resolutions for this crossing (right).

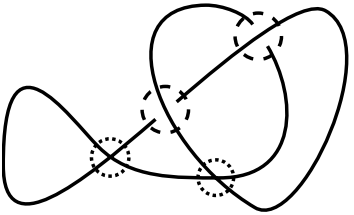

Figure 5. A link pseudodiagram where the upper crossings circled with dashes are resolved and the lower crossings circled with dots are unresolved.

Link equivalence allows the collection of all links to be partitioned into equivalence classes. To determine equivalence on a diagrammatic level, we use *Reidemeister moves* (depicted in Figure 6) and *planar isotopies* (depicted in Figure 7). In this paper, we will focus almost exclusively on the R2 move. Both Reidemeister moves and planar isotopies are local moves, meaning they occur within a fixed region of the plane (so the diagram outside of this region remains unchanged and is, therefore, omitted from Figures 6 and 7). A planar isotopy can be thought of as stretching and bending a single strand of a link diagram without affecting the crossing structure of the diagram. An example of a planar isotopy is shown in Figure 8.

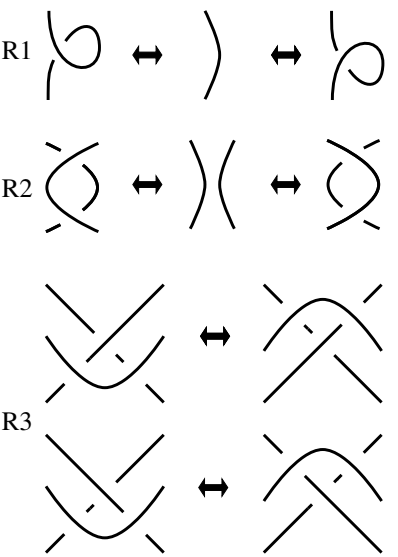

Figure 6. The three Reidemeister moves: R1 moves add or remove a loop, R2 moves overlay one strand on top of a nearby strand (or the reverse process), and R3 moves slide a strand over a crossing. In this paper, we will focus almost exclusively on the R2 move.

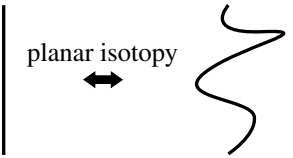

Figure 7. A general planar isotopy.

Definition 2.7. If there is a finite sequence of Reidemeister moves and planar isotopies that turns a given link diagram D_1 into another link diagram D_2, then D_1 and D_2 are called *equivalent*.

As was the case with link equivalence, link diagram equivalence allows the collection of all link diagrams to be partitioned into equivalence classes.

2B. ***Splittable link diagrams, non-self-intersections, and linking numbers.*** The ability to determine whether or not the components of a two-component link diagram can be separated from each other will be crucial for determining the winner of the linking-unlinking game. As such, we divide the family of link diagrams into splittable link diagrams and unsplittable link diagrams.

Definition 2.8. A link diagram is called *splittable* if it is equivalent to a *split link diagram* where one or more components of the diagram can be separated from one or more remaining components of the diagram by a circle. If a link diagram is not splittable (resp. not split), then it is called *unsplittable* (resp. *nonsplit*).

If a link diagram is splittable, we can imagine being able to split one or more components away from the rest of the link diagram. Figures 9 and 10 provide examples of splittable and unsplittable two-component link diagrams, respectively.

Now with an understanding of splittable and unsplittable link diagrams, we will introduce an invariant of two-component links called the linking number which will be used to help us detect when a link diagram is unsplittable. To define this quantity, we first need to discuss orientations of link pseudodiagrams.

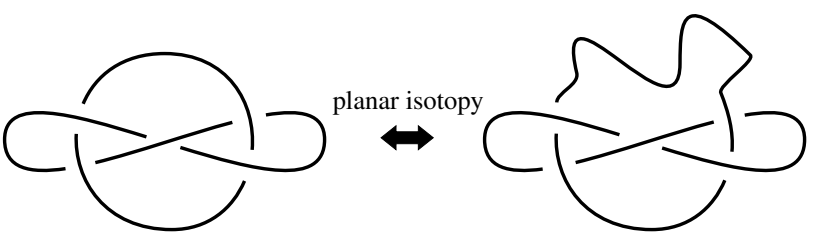

Figure 8. An example of a planar isotopy.

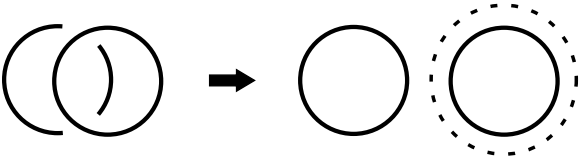

Figure 9. An example of a splittable two-component link diagram (where each link diagram component in this case is an unknotted circle). The dashed circle on the right separates the two unknotted components of the link diagram.

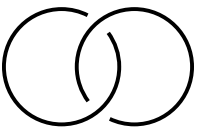

Figure 10. An example of an unsplittable two-component link diagram called the Hopf link diagram.

Definition 2.9. An *oriented* link pseudodiagram is a link pseudodiagram where one of two possible directions of travel has been chosen for each component of the link pseudodiagram.

For an example of an oriented two-component link diagram, see Figure 11. Given an oriented link pseudodiagram, we can associate signs to each resolved crossing.

Definition 2.10. To each resolved crossing of an oriented link pseudodiagram, we can associate a *crossing sign* as shown in Figure 12.

In Figure 11, each crossing is labeled with its sign. The final ingredient needed to define the linking number is the classification of the crossings of a link pseudodiagram into self-intersections and non-self-intersections.

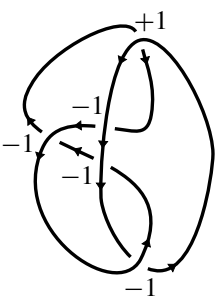

Figure 11. An oriented two-component link diagram labeled with the signs of each of its crossings. Arrows are used to denote the choice of direction of travel.

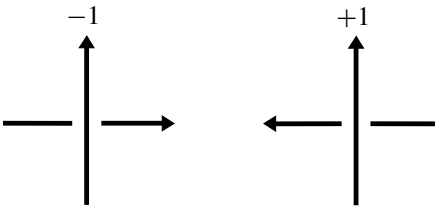

Figure 12. The left figure shows a -1 crossing and the right figure shows a $+1$ crossing.

Definition 2.11. If the two strands meeting at a crossing of a link pseudodiagram come from the same component of the link pseudodiagram, then such a crossing is called a *self-intersection* (*SI*). If a crossing is not a self-intersection, then we call it a *non-self-intersection* (*NSI*).

Note that all crossings of a knot pseudodiagram are necessarily SIs. Consequently, NSIs can only occur in link pseudodiagrams containing at least two components. Figure 13 provides examples of SIs and NSIs in a link diagram. The idea of classifying the crossings of a link pseudodiagram as SIs or NSIs will be used in the definition of the linking number below, as well as later on in this paper.

Definition 2.12. The *linking number* of an oriented two-component link diagram is defined to be half of the sum of the crossing signs, where the sum is taken over all of the non-self-intersections (NSIs) of the link diagram.

Looking back at Figure 11, we see that the two-component link diagram has a linking number of $\frac{1}{2}[3(-1)+1(1)] = -1$. Observe that the leftmost crossing, which has crossing sign -1, is not included in the linking number computation because this crossing is an SI. The linking number is an *invariant* of oriented two-component link diagrams, which means that two equivalent oriented two-component link diagrams have the same linking number. Note that any splittable two-component link diagram has linking number 0, as such a link diagram is equivalent to a link diagram with no NSIs. This tells us, by the contrapositive, that any oriented two-component link

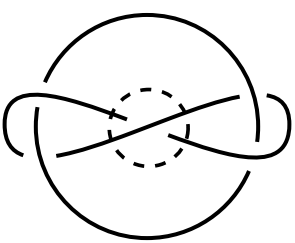

Figure 13. The dashed circle surrounds a crossing that is a self-intersection (SI). All other crossings are non-self-intersections (NSIs).

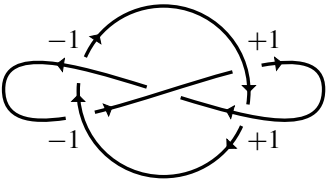

Figure 14. The Whitehead link is unsplittable with a linking number of 0.

diagram with nonzero linking number is unsplittable. An oriented two-component link diagram with linking number 0 is not necessarily splittable, however. An example of this is shown in Figure 14. To compute the linking number in this example, we ignore the center crossing because it is an SI. The two crossings on the left have a negative sign while the two crossings on the right have a positive sign. Summing these signs gives a linking number of $\frac{1}{2}[2(-1)+2(1)]=0$ but the link diagram is known to be unsplittable through other methods.

The following proposition, which will be useful later in this paper, determines the parity of the number of NSIs in a two-component link pseudodiagram.

Proposition 2.13. *Every two-component link pseudodiagram contains an even number of non-self-intersections* (*NSIs*).

Proof. The result is clearly true for a split two-component link pseudodiagram since it contains no NSIs. Now consider a nonsplit two-component link pseudodiagram (which necessarily contains NSIs). Choose a component of the link pseudodiagram and call it component 1. We can distinguish between the "inside" and "outside" of component 1 by giving it a canonical checkerboard coloring, that is, by coloring the regions of the shadow of component 1 either black or white so that regions sharing an edge have opposite colors and so that the unbounded region is colored white. We then view the black regions as the "inside" and the white regions as the "outside" of component 1. An example of a canonical checkerboard coloring is shown in Figure 15. It is a classical result that every link pseudodiagram has a checkerboard coloring. (For more details, see Section X.6 of [Bollobás 1998].)

Call the second component of the link pseudodiagram component 2. Assign component 2 an orientation and choose a starting point on component 2 that is outside of component 1. Since there are NSIs, at some point component 2 will cross from the outside of component 1 to the inside of component 1. Since component 2 enters the inside of component 1, it must also exit the inside of component 1 because component 2 is a closed curve that starts outside of component 1. This follows from the Jordan curve theorem, which states that every simple closed curve in the plane has an interior and an exterior. (For more details, see Chapter 3 of [Henle 1979].) Thus, the link pseudodiagram must contain at least two NSIs.

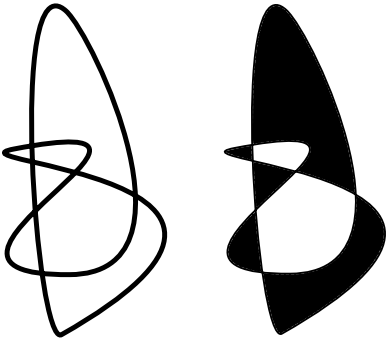

Figure 15. A knot shadow on the left with its canonical checkerboard coloring on the right. The black regions indicate the "inside" and the white regions indicate the "outside" of the knot shadow.

Following the orientation of component 2 from its starting point, the first NSI will bring us from the outside of component 1 to the inside of component 1. The second NSI will bring us back to the outside of component 1. By iterating this argument, we can see that we are outside component 1 after passing through an even number of NSIs and we are inside component 1 after passing through an odd number of NSIs.

Suppose the link pseudodiagram contained an odd number of NSIs. Then, by the argument from the preceding paragraph, we would end up inside component 1 after beginning at the starting point, following the orientation of component 2, and returning to the starting point. But then the endpoint of component 2 is inside component 1 and the starting point of component 2 is outside component 1. This is a contradiction since the starting point and the endpoint of component 2 are the same point. Therefore, there cannot be an odd number of NSIs in a two-component link pseudodiagram. □

2C. ***Rational tangles and rational link diagrams.*** Rational links come from rational tangles that are formed by an iterative process of twisting two strands.

Definition 2.14. We define a *rational tangle*, denoted by $(a_1, a_2, \ldots, a_n)$, where all of the a_i are integers, through the following construction. We begin with two parallel vertical strands, as shown in the top left of Figure 16. We then read $(a_1, a_2, \ldots, a_n)$ from left to right, twisting strands as we go. In particular, we begin by twisting the bottom two endpoints $|a_1|$ times and then proceed to alternate between twisting the right two endpoints $|a_{2k}|$ times and twisting the bottom two endpoints $|a_{2k+1}|$ times. If $a_i > 0$ (resp. $a_i < 0$), then we apply $|a_i|$ twists in such a way that the overstrand has a positive (resp. negative) slope. We call the a_i the *syllables* of the *rational tangle word* $(a_1, a_2, \ldots, a_n)$.

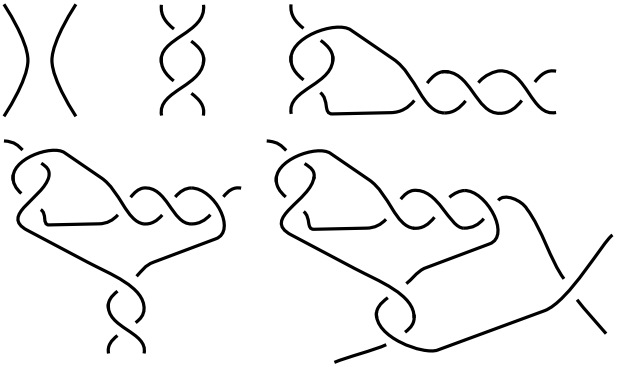

Figure 16. The rational tangles (0) (top left), (2) (top center), (2, −3) (top right), (2, −3, −2) (bottom left), and (2, −3, −2, 1) (bottom right).

See Figure 16 for an example of the construction of a rational tangle. Since the linking-unlinking game will be played on two-component link pseudodiagrams, we need to generalize our notation for rational tangle words to allow for unresolved crossings.

Definition 2.15 [Johnson 2011]. By making a subset of the crossings of a rational tangle unresolved, we create a *rational pseudotangle*. We denote a rational pseudotangle by a *rational pseudotangle word* $(a_1(b_1), \ldots, a_n(b_n))$, where the $a_i(b_i)$ are called the *syllables* of the word and where $|a_i|$ denotes the number of resolved crossings in the i-th syllable and $|b_i|$ denotes the number of unresolved crossings in the i-th syllable. If either $a_i = 0$ or $b_i = 0$ (but not both) for some i, then we omit the single occurrence of 0 or (0) in the rational pseudotangle word. If both $a_i = 0$ and $b_i = 0$ for some i, then we replace $a_i(b_i)$ by 0 in the rational pseudotangle word.

We now present a number of tangle equivalences that will be useful later in this paper.

Proposition 2.16 [Johnson 2011, Lemma 4.2]. *The following statements provide a set of tangle equivalences for rational tangles. Similar statements can also be made for rational pseudotangles*:

(0) $(a_1, \ldots, a_i, 0) = (a_1, \ldots, a_i)$.

(1) $(a_1, \ldots, a_i, 0, a_{i+1}, \ldots, a_n) = (a_1, \ldots, a_i + a_{i+1}, \ldots, a_n)$.

(2) $(a_1, \ldots, a_i, 0, 0, a_{i+1}, \ldots, a_n) = (a_1, \ldots, a_i, a_{i+1}, \ldots, a_n)$.

(3) $(0, a_1 + 1, a_2, \ldots, a_n) = (0, a_1, a_2, \ldots, a_n)$.

(4) $(1, a_1, a_2, \ldots, a_n) = (a_1 + 1, a_2, \ldots, a_n)$.

(5) $(-1, a_1, a_2, \ldots, a_n) = (a_1 - 1, a_2, \ldots, a_n)$.

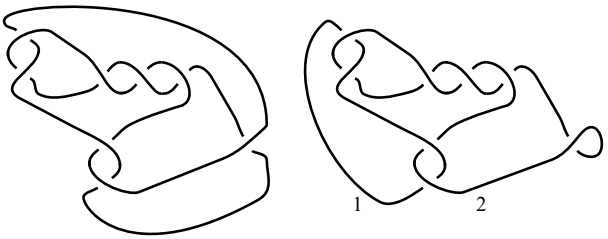

Figure 17. The numerator closure of the tangle $(2, -3, -2, 1)$ produces a knot diagram (left) and the denominator closure of the tangle $(2, -3, -2, 1)$ produces a two-component link diagram where the components have been numbered 1 and 2 (right).

Given that a rational pseudotangle has four endpoints, there are two ways to close a rational pseudotangle to form a rational link pseudodiagram.

Definition 2.17. A rational pseudotangle can be closed to form a *rational link pseudodiagram* either by connecting the top endpoints together and the bottom endpoints together (forming the *numerator closure*) or by connecting the left endpoints together and the right endpoints together (forming the *denominator closure*).

The left side of Figure 17 shows the numerator closure of the rational tangle $(2, -3, -2, 1)$ from Figure 16, which creates a knot diagram, and the right side of Figure 17 shows the denominator closure of this tangle, which creates a two-component link diagram. In general, either both closures of a rational pseudotangle will be knot pseudodiagrams or one closure will be a knot pseudodiagram and the other will be a two-component link pseudodiagram.

2D. ***Non-self-intersections for rational tangles.*** In Definition 2.11, we defined self-intersections (SIs) and non-self-intersections (NSIs) for link pseudodiagrams. We will now define SIs and NSIs for rational pseudotangles.

Definition 2.18. A crossing that occurs between a strand of a rational pseudotangle and itself is called a *self-intersection* (*SI*). Otherwise, the crossing occurs between the two strands and is called a *non-self-intersection* (*NSI*). Furthermore, a syllable of a rational pseudotangle word is called an *SI syllable* (resp. *NSI syllable*) if all of the crossings in the syllable are SIs (resp. NSIs).

If a rational pseudotangle has a closure that is a two-component link pseudodiagram, then the two strands of the pseudotangle become the two separate components of the link pseudodiagram and the SIs and NSIs of the pseudotangle become the SIs and NSIs of the link pseudodiagram, respectively. If a rational pseudotangle does not have a closure that is a two-component link pseudodiagram (if both closures result in a knot pseudodiagram), then both the SIs and the NSIs of the pseudotangle become SIs of the knot pseudodiagram since knot pseudodiagrams cannot contain NSIs.

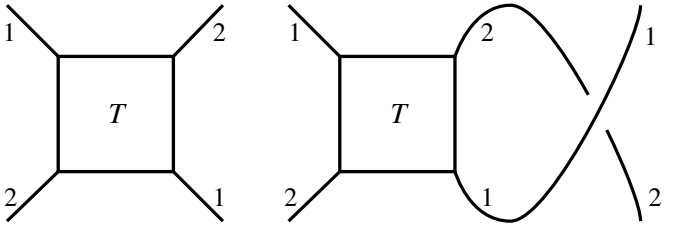

Figure 18. A rational pseudotangle T where strand 1 has endpoints in the northwest and southeast corners and strand 2 has endpoints in the northeast and southwest corners (left) and the result of adding a half-twist to this pseudotangle (right).

When looking for winning strategies for the linking-unlinking game played on rational two-component link shadows, we need to make sure that there actually exists a closure of the rational pseudotangle that is a two-component link pseudodiagram. The following result addresses this issue.

Proposition 2.19. *A rational pseudotangle has a closure that is a two-component link pseudodiagram if and only if the pseudotangle contains an even number of NSIs.*

Proof. ($\Rightarrow$) Assume we have a two-component link pseudodiagram that is a closure of a rational pseudotangle. By Proposition 2.13, the link pseudodiagram must contain an even number of NSIs. This implies that the rational pseudotangle must also contain an even number of NSIs because, otherwise, the rational pseudotangle would contain an odd number of NSIs and closure would create a two-component link pseudodiagram with an odd number of NSIs, contradicting Proposition 2.13.

($\Leftarrow$) Assume we have a rational pseudotangle that contains an even number of NSIs and suppose, for a contradiction, that both the numerator and denominator closures result in a knot pseudodiagram. The only way this can happen is for one strand of the pseudotangle to have endpoints in the northwest and southeast corners and the other strand to have endpoints in the northeast and southwest corners, as shown on the left side of Figure 18.

Add a half-twist to the end of the pseudotangle, as shown on the right side of Figure 18. This produces a rational pseudotangle with an odd total number of NSIs. We can see that the numerator closure of this new rational pseudotangle will result in a two-component link pseudodiagram. But this means we have a two-component link pseudodiagram with an odd total number of NSIs, which contradicts Proposition 2.13. □

In the proposition that follows, we present information about SI syllables and NSI syllables for rational pseudotangle words.

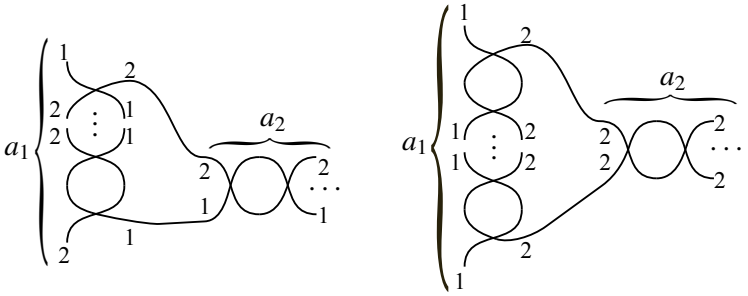

Figure 19. The case where a_1 is odd (left) and the case where a_1 is even (right). The two strands of the pseudotangle are numbered 1 and 2.

Proposition 2.20. *The following statements are true for rational tangle words* $(a_1, \ldots, a_n)$. *Similar statements can also be made for rational pseudotangle words*:

(1) *The first syllable* a_1 *is an NSI syllable.*

(2) *The second syllable* a_2 *is an SI syllable if and only if* a_1 *is even.*

(3) *If* a_i *is an SI syllable, then* a_{i+1} *is an NSI syllable.*

(4) *Assume* a_i *is an SI syllable. Then* a_{i+2} *is an SI syllable if and only if* a_{i+1} *is even.*

(5) *If both* a_i *and* a_{i+1} *are NSI syllables and* a_{i+1} *is odd, then* a_{i+2} *is an SI syllable.*

(6) *If both* a_i *and* a_{i+1} *are NSI syllables and* a_{i+1} *is even, then* a_{i+2} *is an NSI syllable.*

Proof. Condition (1) is true by Definition 2.18 because the crossings of the first syllable a_1 are necessarily between the two strands of the rational tangle. The proof of the forward direction of condition (2) proceeds by contraposition and follows from the left side of Figure 19. The proof of the reverse direction of condition (2) follows from the right side of Figure 19.

For the remainder of this proof, let L denote the rational tangle $(a_1, a_2, \ldots, a_{i-1})$ that precedes the syllable a_i. Note that when we construct a rational tangle, the northwest endpoint of L remains fixed. Label the strand of L incident to this endpoint strand 1 and label the other strand of L strand 2. Furthermore, we can assume that a_i is a horizontal twist without loss of generality, as we can create the case where a_i is a vertical twist by reflecting the diagram over the line $y = -x$.

To prove condition (3), assume a_i is an SI syllable. Then, as shown in Figure 20, the next syllable a_{i+1} will be an NSI syllable. For the remaining three conditions, we will consider cases. Figures 21, 22, and 23 show these six total cases, grouped into three pairs based on the location of the second endpoint of strand 1 of L and

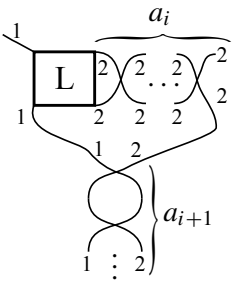

Figure 20. An SI syllable a_i followed by an NSI syllable a_{i+1}.

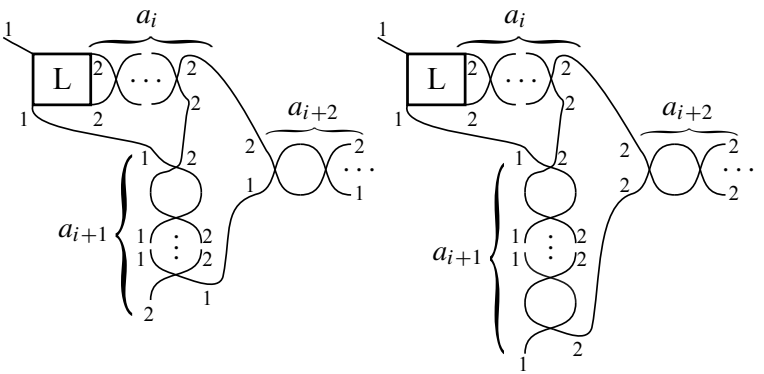

Figure 21. The two cases where the southwest endpoint of L comes from strand 1.

based on the parity of a_{i+1}. To prove condition (4), assume a_i is an SI syllable. The proof of the forward direction of condition (4) proceeds by contraposition and follows from the left side of Figure 21. The proof of the reverse direction of condition (4) follows from the right side of Figure 21. To prove conditions (5) and (6) (combined), assume both a_i and a_{i+1} are NSI syllables. The proof of these conditions follows from Figures 22 and 23.

□

We now present notation used to highlight the presence of isolated SI syllables in rational tangle words.

Notation 2.21. Given a syllable a of a rational tangle word, we use a^* to indicate that the syllable a is an SI syllable.

Given Proposition 2.20, we now present a result that leads to a method to decompose any rational pseudotangle word into SI syllables and strings of NSI syllables.

Proposition 2.22. *Every rational tangle word can be decomposed into strings of NSI syllables that alternate with isolated SI syllables. Furthermore*:

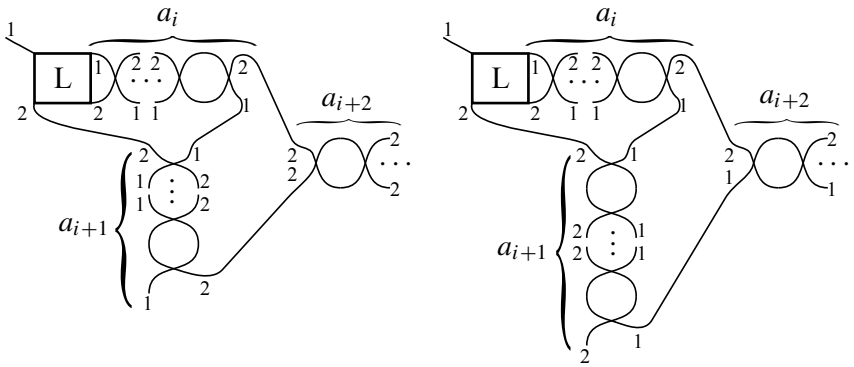

Figure 22. The two cases where the northeast endpoint of L comes from strand 1 and where a_i and a_{i+1} are NSI syllables. Note that a_i must be odd for a_{i+1} to be an NSI syllable.

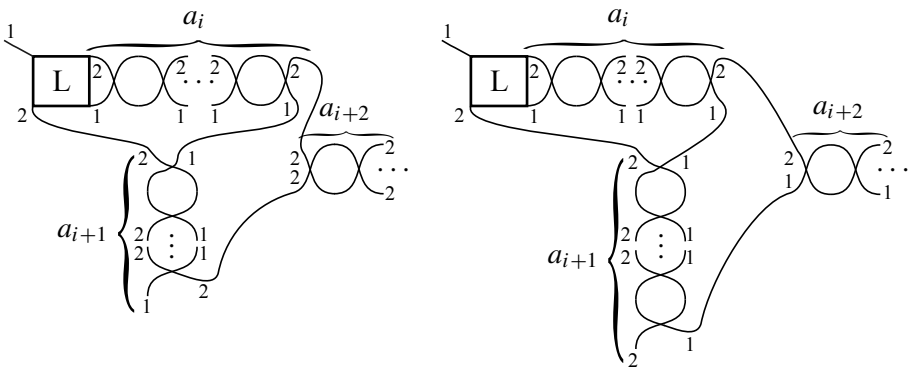

Figure 23. The two cases where the southeast endpoint of L comes from strand 1 and where a_i and a_{i+1} are NSI syllables. Note that a_i must be even for a_{i+1} to be an NSI syllable.

(1) *All but the last string of NSI syllables consists of either*
- *a single even syllable*,
- *two consecutive odd syllables*, *or*
- *an odd syllable followed by an arbitrary nonempty string of even syllables followed by a final odd syllable.*

(2) *If the rational tangle contains an even number of NSIs*, *then the last string of NSI syllables consists of the same possibilities as condition* (1).

(3) *If the rational tangle contains an odd number of NSIs*, *then the last string of NSI syllables consists of either*
- *a single odd syllable*, *or*
- *an odd syllable followed by an arbitrary nonempty string of even syllables.*

A similar statement can also be made for rational pseudotangle words.

Proof. By Proposition 2.20(3), no two SI syllables can be adjacent. Thus, the strings of NSI syllables alternate with isolated SI syllables.

We will now prove condition (1). Suppose we have a nonfinal string of NSI syllables. We want to show that this string consists of a single even syllable or an odd syllable followed by an arbitrary (possibly empty) string of even syllables followed by a final odd syllable. In Cases 1 and 2 below, we consider the first nonfinal string of NSI syllables, whose first syllable is a_1. By Proposition 2.20(1), we know that a_1 is an NSI syllable.

Case 1: Suppose a_1 is even. Then Proposition 2.20(2) implies that a_2 is an SI syllable, so the first nonfinal string of NSI syllables consists of the single even syllable a_1.

Case 2: Suppose a_1 is odd. Then Proposition 2.20(2) implies that a_2 is an NSI syllable. If a_2 is odd, then Proposition 2.20(5) implies that a_3 is an SI syllable, which gives a string of NSI syllables composed of two odd syllables. If a_2 is even, then Proposition 2.20(6) implies that a_3 is an NSI syllable. This string of even NSI syllables continues by repeatedly applying Proposition 2.20(6). This string of even NSI syllables must eventually terminate, however, since this is a nonfinal string of NSI syllables. Therefore, we eventually find an odd syllable in this string of NSI syllables; call this syllable a_k. This syllable a_k is the last syllable in this string of NSI syllables because the next syllable a_{k+1} is an SI syllable by Proposition 2.20(5). Thus, the first nonfinal string of NSI syllables consists of an odd syllable a_1 followed by an arbitrary (possibly empty) string $a_2, \ldots, a_{k-1}$ of even syllables followed by a final odd syllable a_k.

In Cases 3 and 4 below, we consider a nonfirst nonfinal string of NSI syllables. Let a_i, where $i > 1$, denote the first syllable in this string of NSI syllables.

Case 3: Suppose a_i is even. Since a_i is the first NSI syllable in a nonfirst string of NSI syllables, we know that a_{i-1} is an SI syllable. By Proposition 2.20(4), a_{i+1} is an SI syllable, so the nonfirst nonfinal string of NSI syllables consists of the single even syllable a_i.

Case 4: Suppose a_i is odd. Since a_i is the first NSI syllable in a nonfirst string of NSI syllables, we know that a_{i-1} is an SI syllable. Then Proposition 2.20(4) implies that a_{i+1} is an NSI syllable. The remainder of this case is similar to Case 2, except that a_{i+1} is now playing the role of a_2.

This completes the proof of condition (1). To prove condition (2) and condition (3), we will consider the final string of NSI syllables. Let a_i be the first syllable in this final string of NSI syllables. We will consider two cases based on the parity of the NSIs.

To prove condition (2), suppose there are an even number of NSIs in the tangle word. Condition (1) of this proposition implies that each nonfinal string of NSI syllables contains an even number of NSIs. Thus, the final string of NSI syllables must contain an even number of NSIs for the tangle word to contain an even total number of NSIs.

Case A: Suppose a_i is even. If the final string of NSI syllables consists of the syllable a_i alone, then we have the desired result. We may now assume that the final string of NSI syllables contains a second syllable a_{i+1}. If the final string of NSI syllables is the only string of NSI syllables in the tangle word, then $a_i = a_1$, $a_{i+1} = a_2$, and the argument from Case 1 gives the desired result. Now suppose there are at least two strings of NSI syllables in the tangle word. Then the argument from Case 3 gives the desired result.

Case B: Suppose a_i is odd. If the final string of NSI syllables consists of the syllable a_i alone, then we have a contradiction of the fact that the final string of NSI syllables contains an even number of NSIs. We may now assume that the final string of NSI syllables contains a second syllable a_{i+1}.

Subcase 1: Suppose the final string of NSI syllables is the only string of NSI syllables in the tangle word. Then $a_i = a_1$ and $a_{i+1} = a_2$.

Suppose a_2 is odd. If the final string of NSI syllables contains only a_1 and a_2, then we have the desired result. If the final string of NSI syllables contains a third syllable a_3, then the argument from the first three sentences of Case 2 gives the desired result.

Suppose a_2 is even. If the final string of NSI syllables contains only a_1 and a_2, then we have a contradiction of the fact that the final string of NSI syllables contains an even number of NSIs. If the final string of NSI syllables contains a third syllable a_3, then the argument from Case 2 implies that a_1 is followed by a nonempty string of even syllables. Eventually, an odd syllable must occur because the total number of NSIs in the final string of NSI syllables must be even. Then either this odd syllable is the last syllable of the tangle word or the next syllable is an SI syllable by Proposition 2.20(5). In either case, we have the desired result.

Subcase 2: Now suppose there are at least two strings of NSI syllables in the tangle word.

Suppose a_{i+1} is odd. If the final string of NSI syllables contains only a_i and a_{i+1}, then we have the desired result. If the final string of NSI syllables contains a third syllable a_{i+2}, then the argument from Case 4 gives the desired result.

Suppose a_{i+1} is even. If the final string of NSI syllables contains only a_i and a_{i+1}, then we have a contradiction of the fact that the final string of NSI syllables contains an even number of NSIs. If the final string of NSI syllables contains a third syllable a_{i+2}, then the argument from Case 4 implies that a_i is followed by a

nonempty string of even syllables. Eventually, an odd syllable must occur because the total number of NSIs in the final string of NSI syllables must be even. Then either this odd syllable is the last syllable of the tangle word or the next syllable is an SI syllable by Proposition 2.20(5). In either case, we have the desired result.

To prove condition (3), suppose there are an odd number of NSIs in the tangle word. Condition (1) of this proposition implies that each nonfinal string of NSI syllables contains an even number of NSIs. Thus, the final string of NSI syllables must contain an odd number of NSIs for the tangle word to contain an odd total number of NSIs.

Case A: Suppose a_i is even. If the final string of NSI syllables consists of the syllable a_i alone, then we have a contradiction of the fact that the final string of NSI syllables contains an odd number of NSIs. We may now assume that the final string of NSI syllables contains a second syllable a_{i+1}. If the final string of NSI syllables is the only string of NSI syllables in the tangle word, then $a_i = a_1$, $a_{i+1} = a_2$, and the argument from Case 1 gives a contradiction of the fact that the final string of NSI syllables contains an odd number of NSIs. Now suppose there are at least two strings of NSI syllables in the tangle word. Then the argument from Case 3 gives a contradiction of the fact that the final string of NSI syllables contains an odd number of NSIs.

Case B: Suppose a_i is odd. If the final string of NSI syllables consists of the syllable a_i alone, then we have the desired result. We may now assume that the final string of NSI syllables contains a second syllable a_{i+1}.

Subcase 1: Suppose the final string of NSI syllables is the only string of NSI syllables in the tangle word. Then $a_i = a_1$ and $a_{i+1} = a_2$.

Suppose a_2 is odd. If the final string of NSI syllables contains only a_1 and a_2, then we have a contradiction of the fact that the final string of NSI syllables contains an odd number of NSIs. If the final string of NSI syllables contains a third syllable a_3, then the argument from the first three sentences of Case 2 gives a contradiction of the fact that the final string of NSI syllables contains an odd number of NSIs.

Suppose a_2 is even. If the final string of NSI syllables contains only a_1 and a_2, then we have the desired result. If the final string of NSI syllables contains a third syllable a_3, then the argument from Case 2 implies that a_1 is followed by a nonempty string of even syllables which must terminate because the tangle word is finite. This gives the desired result.

Subcase 2: Now suppose there are at least two strings of NSI syllables in the tangle word.

Suppose a_{i+1} is odd. If the final string of NSI syllables contains only a_i and a_{i+1}, then we have a contradiction of the fact that the final string of NSI syllables contains an odd number of NSIs. If the final string of NSI syllables contains a third

syllable a_{i+2}, then the argument from Case 4 gives a contradiction of the fact that the final string of NSI syllables contains an odd number of NSIs.

Suppose a_{i+1} is even. If the final string of NSI syllables contains only a_i and a_{i+1}, then we have the desired result. If the final string of NSI syllables contains a third syllable a_{i+2}, then the argument from Case 4 implies that a_i is followed by a nonempty string of even syllables which must terminate because the tangle word is finite. This gives the desired result. □

We now present an example of applying Proposition 2.22 to decompose a rational tangle word into SI syllables and NSI syllables.

Example 2.23. Consider the rational tangle word

$$(1, 4, 2, 1, 3, 5, 3, 2, 1, 2, 0, 5, 2, 6, 4).$$

Reading from left to right, we see that our first string of NSI syllables is 1, 4, 2, 1, as this is an odd syllable followed by a nonempty string of even syllables followed by a final odd syllable. The next syllable, 3, is an SI syllable. We now update our tangle word to get

$$(1, 4, 2, 1, 3^*, 5, 3, 2, 1, 2, 0, 5, 2, 6, 4).$$

Proceeding to the right, we see that 5, 3 is the next string of NSI syllables, as this is a string of two consecutive odd syllables. The next syllable, 2, is an SI syllable, so we update our tangle word to get

$$(1, 4, 2, 1, 3^*, 5, 3, 2^*, 1, 2, 0, 5, 2, 6, 4).$$

Continuing this process, we see that 1, 2, 0, 5 is the next string of NSI syllables and the following syllable, 2, is an SI syllable. Updating our tangle word, we get

$$(1, 4, 2, 1, 3^*, 5, 3, 2^*, 1, 2, 0, 5, 2^*, 6, 4).$$

The final string of NSI syllables consists of the single even syllable 6, so our rational tangle word ends with the syllable 4, which is an SI syllable. Thus, our completely decomposed tangle word is given by

$$(1, 4, 2, 1, 3^*, 5, 3, 2^*, 1, 2, 0, 5, 2^*, 6, 4^*).$$

3. The linking-unlinking game

In this section, we will define the linking-unlinking game and provide winning strategies for the linking-unlinking game played on all shadows of two-component rational tangle closures and played on a large family of general two-component link shadows.

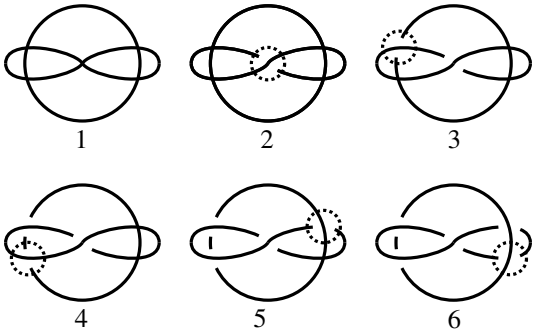

Figure 24. The linking-unlinking game played on a shadow of the Whitehead link. Crossings being played on are surrounded by a dotted circle.

3A. ***Defining the linking-unlinking game.*** In what follows, we introduce the linking-unlinking game and present two key player strategies that will be used often in the remainder of this paper.

Definition 3.1. The *linking-unlinking game* is a two-player game played on the shadow of a two-component link. The game is played with each player taking turns resolving crossings of the link shadow. The game ends when all of the crossings are resolved and a two-component link diagram is formed. One player, the *linker*, wins if the resulting link diagram is unsplittable. The other player, the *unlinker*, wins if the resulting link diagram is splittable.

An example of the linking-unlinking game is provided below.

Example 3.2. Figure 24 shows an example of the linking-unlinking game being played on a shadow of the Whitehead link. Here, the unlinker goes first. The second frame shows that the unlinker decides to play on the central crossing. Next, the linker plays on the upper left crossing, as shown in the third frame. In the fourth frame, the unlinker responds on the lower left crossing, resolving the crossing with the intention of being able to reduce the two left crossings with an R2 move. Next, the linker plays on the top right crossing, as shown in the fifth frame. Finally, the sixth frame shows that the unlinker responds on the lower right crossing, resolving the crossing with the intention of being able to reduce the two right crossings with an R2 move. The link diagram at the end of the game is equivalent to the splittable trivial two-component link diagram (also called the two-component unlink diagram), so the unlinker wins. Figure 25 shows this equivalence.

Given the definition of the linking-unlinking game as a two-player combinatorial game, our main goal will now be to determine who wins the game and who loses the game for various two-component link shadows.

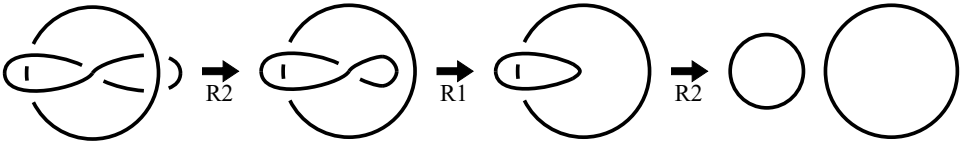

Figure 25. A sequence of Reidemeister moves used to show that the two-component link diagram above on the left is splittable.

Definition 3.3. In a two-player game, a player has a *winning strategy* if there is always a sequence of moves that allows them to win, regardless of what their opponent does.

Given the structure of rational pseudotangles, we will often group together crossings in a syllable of the pseudotangle word when constructing winning strategies for the linking-unlinking game.

Definition 3.4. A *subtwist region* in a link pseudodiagram is a section of the associated link shadow consisting of a chain of bigons. A *twist region* in a link pseudodiagram is a section of the associated link shadow consisting of a maximal chain of bigons, maximal in the sense that the chain cannot be extended by adding more bigons. Note that a twist region consisting of a single crossing with no incident bigons is possible.

A chain of bigons can be thought of as being formed by twisting two parallel strands. Figure 26 provides a schematic depiction of a twist region. Given a subtwist region of a link pseudodiagram, we now define an operation called flyping that will be useful in defining two important player strategies for the linking-unlinking game.

Definition 3.5. A *flype* is a move performed on a tangle of a link pseudodiagram that reflects the tangle over an axis.

See Figure 27 for an example of a flype. For our purposes, we will only consider flypes applied to subtwist regions of a tangle pseudodiagram. In what follows, we present two key player strategies for the linking-unlinking game that will be applied

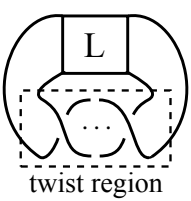

Figure 26. A schematic diagram where the twist region is surrounded by a dashed rectangle. The box labeled L represents the remainder of the link pseudodiagram.

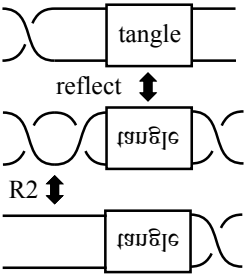

Figure 27. A schematic diagram of a flype applied to a tangle in a link pseudodiagram.

throughout the remainder of this paper. Note that these are response strategies for the next player regardless of the role of the current player.

Strategy 3.6 (R2-strategy for the next player). Assign the given link pseudodiagram an orientation. When a player plays by resolving a crossing in a twist region (consequently assigning this crossing a crossing sign), the next player responds by resolving a crossing in the same twist region so that this crossing has crossing sign opposite to the previous resolved crossing. This allows both crossings to be removed by applying a flype and an R2 move, as shown in Figure 28.

The R2-strategy is useful for both players because it allows for pairs of crossings to be removed at the end of the game.

Strategy 3.7 (anti-R2-strategy for the next player). Assign the given link pseudodiagram an orientation. Here, when a player plays by resolving a crossing in a twist region, the next player responds by resolving a crossing in the same twist region

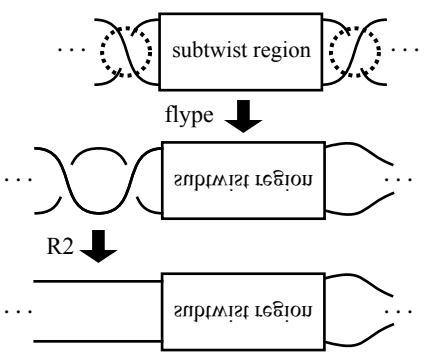

Figure 28. An example of the R2-strategy. A player resolves a crossing and the next player responds on another crossing in the twist region so that a flype and an R2 move will remove the pair of crossings.

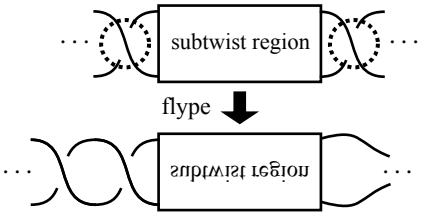

Figure 29. An example of the anti-R2-strategy. A player resolves a crossing and the next player responds on another crossing in the twist region so that a flype will create a clasp between the two strands of the twist region.

so that this crossing has the same crossing sign as the previous resolved crossing. In this situation, the two crossings cannot be removed by applying a flype to the subtwist region and one of the two incident crossings, as shown in Figure 29. Note that one or both of these crossings may be able to be removed, however, by applying Reidemeister moves to the remainder of the link diagram at the end of the game.

The anti-R2-strategy is particularly useful for the linker because it creates a clasp between two strands. As an example of why creating clasps might be useful for the linker, consider the unsplittable Hopf link diagram in Figure 10.

3B. ***The linking-unlinking game for rational two-component link shadows.*** In this section, we provide winning strategies for the linking-unlinking game played on shadows of two-component rational tangle closures. We begin by providing a lemma that will be used in the proofs of a number of theorems in the remainder of this paper.

Lemma 3.8. *The following statements are true*:

(1) *The rational tangle* (0) *has a splittable two-component link closure.*

(2) *The rational tangle* (± 2) *has an unsplittable two-component link closure.*

Proof. To prove statement (1), observe from the top left of Figure 16 that the denominator closure of the tangle (0) gives the trivial two-component link diagram. This link diagram is clearly splittable.

To prove statement (2), observe that the denominator closure of the tangle (± 2) gives either the Hopf link diagram from Figure 10 or its reflection. In either case, after orienting this link diagram, it can be seen that the linking number of the resulting oriented two-component link diagram has absolute value 1. Since this link diagram has nonzero linking number, it is unsplittable. □

We now begin our study of the linking-unlinking game by focusing on certain families of rational two-component link shadows. The following result presents winning strategies for the unlinker in a very specific pathological case and for the second player in other cases.

Theorem 3.9. *Suppose we have a shadow of a rational two-component link coming from a closure of the rational tangle* $(a_1, \ldots, a_n)$:

(1) *If* $a_{2k+1} = 0$ *for all* k, *then the unlinker wins.*

(2) *If* $a_{2k+1} \neq 0$ *for at least one* k *and all of the* a_i *are even, then the second player has a winning strategy.*

Proof. We begin by proving statement (1). Assume we have a shadow of a rational two-component link coming from a closure of the rational tangle $(a_1, \ldots, a_n)$ where $a_{2k+1} = 0$ for all k. If $n = 1$, then the tangle word is (0) and the unlinker wins by Lemma 3.8. Now suppose that $n \geq 2$. Then the tangle word can be written in the form $(0, a_2, \ldots, 0, a_{2m})$ or $(0, a_2, \ldots, 0, a_{2m}, 0)$. By repeatedly applying Proposition 2.16(3), we can reduce each of the tangle subwords $(0, a_{2k})$ to $(0, 0)$. By repeatedly applying Proposition 2.16(2), we can remove all but the last copy of $(0, 0)$ from the tangle word. At the end of this process, what remains is either $(0, 0)$ or $(0, 0, 0)$, both of which are equivalent to (0) by applying Proposition 2.16(0) either once or twice, respectively. By Lemma 3.8, the unlinker wins.

We will now prove statement (2). Assume we have a shadow of a rational two-component link coming from a closure of the rational tangle $(a_1, \ldots, a_n)$ where $a_{2k+1} \neq 0$ for at least one k and where all of the a_i are even. Fix a value j such that $a_{2j+1} \neq 0$. We consider two cases, depending on the role of the second player.

Case 1: Suppose the unlinker plays second. Whenever the linker plays on a syllable, the unlinker should respond on the same syllable by using the R2-strategy. Note that this response strategy is always possible since all of the a_i are even. Thus, each time the unlinker responds, two crossings can be removed at the end of the game. After applying flypes and R2 moves at the end of the game, the resulting tangle word will be of the form $(0, \ldots, 0)$. By repeatedly applying Proposition 2.16(0), this tangle word is equivalent to (0) and the unlinker wins by Lemma 3.8.

Case 2: Suppose the linker plays second. Whenever the unlinker plays on any a_i-syllable where $i \neq 2j + 1$, the linker should respond on this syllable by using the R2-strategy. Again, this strategy can be applied since each a_i is even. Thus, each time the linker responds, two crossings can be removed at the end of the game. When the unlinker plays on $a_{2j+1} \neq 0$, the linker should respond by using the R2-strategy until only two unresolved crossings remain in this subtwist region. On the last two unresolved crossings in this subtwist region, the linker should respond by using the anti-R2-strategy.

After applying flypes and R2 moves at the end of the game, the resulting tangle word will be of the form $(0, \ldots, 0, 2, 0, \ldots, 0)$ or $(0, \ldots, 0, -2, 0, \ldots, 0)$, where there are an even number of zeros before the 2 or -2. By repeatedly applying Proposition 2.16(2), we can reduce the tangle word to either $(2, 0, \ldots, 0)$ or

$(-2, 0, \ldots, 0)$. By repeatedly applying Proposition 2.16(0), these tangle words are equivalent to either (2) or (-2), respectively, and the linker wins by Lemma 3.8. □

We now present winning strategies for rational two-component link shadows whose tangle word consists of two odd syllables.

Theorem 3.10. *Playing on the shadow of a rational two-component link coming from a closure of the rational tangle* (a_1, a_2), *if* a_1 *and* a_2 *are both odd, then the second player has a winning strategy.*

Proof. Assume we have a shadow of a rational two-component link coming from a closure of the rational tangle (a_1, a_2) where a_1 and a_2 are both odd. We will consider two cases, depending on the role of the second player.

Case 1: Suppose the linker plays second. When the unlinker plays, the linker should respond on the same syllable using the R2-strategy so long as this is possible. Since the unlinker is playing first and since both of the syllables are odd, the unlinker will necessarily have to resolve the last crossing in either a_1 or a_2. Once the unlinker fully resolves an a_i-syllable, the pseudotangle word can be reduced to one of $(1, (a))$, $(-1, (a))$, $((a), 1)$, or $((a), -1)$, where a is odd. The linker should then play on a by resolving a crossing to have the same overcrossing slope as the previous crossing resolved by the unlinker. Thus, the linker will turn $(1, (a))$ into $(1, 1(|a| - 1))$, $(-1, (a))$ into $(-1, -1(|a| - 1))$, $((a), 1)$ into $(1(|a| - 1), 1)$, or $((a), -1)$ into $(-1(|a| - 1), -1)$.

Since $|a| - 1$ is even, the linker should resume using the R2-strategy to respond to the unlinker until the end of the game. This results in a rational tangle word that can be reduced to either $(1, 1)$ or $(-1, -1)$, which is equivalent to either (2) or (-2) by applying statement (4) or (5) of Proposition 2.16, respectively. By Lemma 3.8, the linker wins.

Case 2: Suppose the unlinker plays second. The unlinker's strategy will be similar to the linker's strategy from Case 1. First, the unlinker should use the R2-strategy until the linker fully resolves an a_i-syllable. As in the previous case, the pseudotangle word can be reduced to the form $(1, (a))$, $(-1, (a))$, $((a), 1)$, or $((a), -1)$, where a is odd. The unlinker should then play on a by resolving a crossing to have overcrossing slope opposite to the previous crossing resolved by the linker. Thus, the unlinker will turn $(1, (a))$ into $(1, -1(|a| - 1))$, $(-1, (a))$ into $(-1, 1(|a| - 1))$, $((a), 1)$ into $(-1(|a| - 1), 1)$, or $((a), -1)$ into $(1(|a| - 1), -1)$.

Since $|a| - 1$ is even, the unlinker should resume using the R2-strategy to respond to the linker until the end of the game. This results in a rational tangle word that can be reduced to either $(1, -1)$ or $(-1, 1)$, both of which are equivalent to (0) by applying statement (4) or (5) of Proposition 2.16, respectively. By Lemma 3.8, the unlinker wins. □

The following theorem builds on the results of the previous two theorems.

Theorem 3.11. *Playing on the shadow of a rational two-component link coming from a closure of the rational tangle* $(a_1, \ldots, a_n)$ *where* $n \geq 3$, *if* a_1 *and* a_n *are odd and all other* a_i *are even, then the second player has a winning strategy.*

Proof. Assume we have a shadow of a rational two-component link coming from a closure of the rational tangle $(a_1, \ldots, a_n)$ where $n \geq 3$, where a_1 and a_n are odd, and where all other a_i are even. When the first player plays on a_1 or a_n, the second player should respond by using the same strategy as used in the proof of Theorem 3.10, as though the game were played on (a_1, a_n), where the linker wants to reduce (a_1, a_n) to either $(1, 1)$ or $(-1, -1)$ and the unlinker wants to reduce (a_1, a_n) to either $(1, -1)$ or $(-1, 1)$. When the first player plays on one of the even a_i-syllables, the second player should respond on the same syllable by using the R2-strategy. At the end of the game, there are two cases for the resulting tangle word, depending on the role of the second player.

Case 1: Suppose the linker plays second. Then the resulting tangle word can be reduced to either $(1, 0, \ldots, 0, 1)$ or $(-1, 0, \ldots, 0, -1)$. If there is an even number of zeros, then the tangle word can be reduced to either $(1, 1)$ or $(-1, -1)$ by repeatedly applying Proposition 2.16(2) and reduced to either (2) or (-2) by applying statement (4) or (5) of Proposition 2.16, respectively. If there is an odd number of zeros, then the tangle word can be reduced to either $(1, 0, 1)$ or $(-1, 0, -1)$ by repeatedly applying Proposition 2.16(2) and reduced to either (2) or (-2) by applying Proposition 2.16(1). By Lemma 3.8, the linker wins.

Case 2: Suppose the unlinker plays second. Then the resulting tangle word can be reduced to either $(1, 0, \ldots, 0, -1)$ or $(-1, 0, \ldots, 0, 1)$. If there is an even number of zeros, then the tangle word can be reduced to either $(1, -1)$ or $(-1, 1)$ by repeatedly applying Proposition 2.16(2) and reduced to (0) by applying statement (4) or (5) of Proposition 2.16, respectively. If there is an odd number of zeros, then the tangle word can be reduced to either $(1, 0, -1)$ or $(-1, 0, 1)$ by repeatedly applying Proposition 2.16(2) and reduced to (0) by applying Proposition 2.16(1). By Lemma 3.8, the unlinker wins. □

We will now explore how the parity of the number of SIs in the rational two-component link shadow affects which player has a winning strategy. The following key theorem, when paired with Theorem 3.9(1), provides winning strategies for all shadows of two-component rational tangle closures. Its proof combines a number of results from earlier in this paper.

Theorem 3.12. *Suppose we have a shadow of a rational two-component link coming from a closure of the rational tangle* $(a_1, \ldots, a_n)$. *Furthermore, assume* $a_{2k+1} \neq 0$ *for some* k:

(1) *If the tangle word contains an even number of SIs, then the second player has a winning strategy.*

(2) *If the tangle word contains an odd number of SIs, then the first player has a winning strategy.*

Proof. Assume we have a shadow of a rational two-component link coming from a closure of the rational tangle $(a_1, \ldots, a_n)$ where $a_{2k+1} \neq 0$ for some k.

We begin by proving statement (1). Assume $(a_1, \ldots, a_n)$ contains an even number of SIs. Using Proposition 2.22, we can decompose $(a_1, \ldots, a_n)$ into alternating strings of NSI syllables and isolated SI syllables. Whenever the first player plays on an SI, the second player should respond by playing arbitrarily on any other SI. Note that this response strategy is always possible since the number of SIs is even.

By Proposition 2.19, since we have a two-component link shadow coming from a closure of a rational tangle, the tangle must contain an even number of NSIs. Therefore, by Proposition 2.22, we have that a nonfinal or final string of NSI syllables can be either (a) a single even syllable, (b) two consecutive odd syllables, or (c) an odd syllable followed by an arbitrary nonempty string of even syllables followed by a final odd syllable. Notice that the proofs of Theorems 3.9(2), 3.10, and 3.11 provide the second player with a strategy for playing on the respective type (a), type (b), and type (c) strings of NSI syllables listed above.

When the first player plays on an unresolved crossing from a nonfinal string of NSI syllables, the second player should respond on an unresolved crossing in the same nonfinal string of NSI syllables by using the strategy for when the unlinker plays second provided by the proof of the theorem corresponding to the type of string of NSI syllables. This strategy ensures that each nonfinal string of NSI syllables can be reduced to the tangle subword (0).

When the first player plays on an unresolved crossing from the final string of NSI syllables, the second player should respond on an unresolved crossing in the same final string of NSI syllables by using the strategy that corresponds to their role as linker or unlinker from the proof of the theorem corresponding to the type of string of NSI syllables.

Note that the second player will always be able to respond on the same nonfinal or final string of NSI syllables because each such string contains an even total number of crossings. We will now consider two cases, depending on the role of the second player.

Case 1: Suppose the unlinker plays second. Let $(a_1, \ldots, a_k)$ denote the first string of NSI syllables. By the proofs of Theorems 3.9(2), 3.10, and 3.11, we know that the unlinker's strategy results in this string of syllables being able to be reduced to the tangle subword (0).

Subcase 1: Suppose the entire tangle word contains a single string of NSI syllables. If the entire tangle word consists of a single string of NSI syllables without an SI syllable at the end of the tangle word, then the tangle word can be reduced to (0) and the unlinker wins by Lemma 3.8. If the entire tangle word consists of a single string of NSI syllables followed by an SI syllable, call it a^*, at the end of the tangle word, then the tangle word can be reduced to $(0, a^*)$, which can be reduced to $(0, 0)$ by repeatedly applying Proposition 2.16(3). By applying Proposition 2.16(0), $(0, 0)$ can be reduced to (0) and the unlinker wins by Lemma 3.8.

Subcase 2: Suppose the tangle word contains multiple strings of NSI syllables. Then we can apply the strategy from all but the last sentence of Subcase 1 to each tangle subword $(a_{k+1}, a_{k+2}, \ldots, a_m, b_m^*)$, where $(a_{k+1}, a_{k+2}, \ldots, a_m)$ is a nonfinal string of NSI syllables and where b_m^* is an SI syllable, so that each such subword can be reduced to $(0, 0)$. These subwords can then be removed from the tangle word by applying Proposition 2.16(2). If the final string of NSI syllables is not followed by a final SI syllable, then the tangle word can be reduced to (0) and the unlinker wins by Lemma 3.8. If the final string of NSI syllables is followed by a final SI syllable, then the argument from Subcase 1 can be applied in its entirety to conclude that the unlinker wins by Lemma 3.8.

Case 2: Suppose the linker plays second. Then, as indicated in the paragraphs preceding Case 1, the only difference in strategy will occur in the final string of NSI syllables. Specifically, the linker (and the unlinker) will respond so that all of the nonfinal strings of NSI syllables can be reduced to the tangle subword (0) and the linker will respond so that the final string of NSI syllables can be reduced to either the tangle subword (2) or the tangle subword (-2).

Subcase 1: Suppose the entire tangle word contains a single string of NSI syllables. If the entire tangle word consists of a single string of NSI syllables without an SI syllable at the end of the tangle word, then the tangle word can be reduced to (2) or (-2) and the linker wins by Lemma 3.8. If the entire tangle word consists of a single string of NSI syllables followed by an SI syllable, call it a^*, at the end of the tangle word, then the tangle word can be reduced to $(2, a^*)$ or $(-2, a^*)$. Since SI syllables do not contribute to the linking number, an argument similar to the one from the proof of Lemma 3.8(2) can be used to conclude that the linker wins.

Subcase 2: Suppose the tangle word contains multiple strings of NSI syllables. Then we can apply the strategy from all but the last sentence of Subcase 1 of Case 1 to each tangle subword $(a_{k+1}, a_{k+2}, \ldots, a_m, b_m^*)$, where $(a_{k+1}, a_{k+2}, \ldots, a_m)$ is a nonfinal string of NSI syllables and where b_m^* is an SI syllable, so that each such subword can be reduced to $(0, 0)$. These subwords can then be removed from the tangle word by applying Proposition 2.16(2). If the final string of NSI syllables is not followed by a final SI syllable, then the tangle word can be reduced to (2)

or (-2) and the linker wins by Lemma 3.8. If the final string of NSI syllables is followed by a final SI syllable, then the argument from Subcase 1 can be applied in its entirety to conclude that the linker wins by Lemma 3.8.

We will now prove statement (2). Assume $(a_1, \ldots, a_n)$ contains an odd number of SIs. Then the first player should begin by playing arbitrarily on an SI. This will effectively reduce the game to one with an even number of unresolved SIs where the first player is now playing second on this new link pseudodiagram. But then the first player can use the strategy from the proof of statement (1) to guarantee a win. □

3C. ***The linking-unlinking game for general two-component link shadows.*** We now broaden our focus from rational two-component link shadows to general two-component link shadows. In particular, we will focus the majority of our attention on finding winning strategies for the linker. First, we introduce the notion of a partial linking number that can be applied to a link pseudodiagram throughout the linking-unlinking game.

Definition 3.13. The *pseudolinking number* of an oriented two-component link pseudodiagram is defined to be half of the sum of the crossing signs, where the sum is taken over all of the resolved non-self-intersections (NSIs) of the link pseudodiagram. Note that an oriented two-component link shadow has a pseudolinking number of 0 and note that the pseudolinking number of an oriented two-component link diagram is the linking number of the diagram.

In what follows, we use the pseudolinking number, the presence of NSIs, and the parity of the number of SIs to provide winning strategies for the linking-unlinking game played on a large family of general two-component link shadows. In particular, we present winning strategies for the unlinker in a very specific pathological case and for the linker in other cases.

Theorem 3.14. *Suppose we have a shadow of a two-component link*:

(1) *If the shadow contains no NSIs, then the unlinker wins.*

(2) *If the shadow contains a nonzero number of NSIs and an even number of SIs, then the linker has a winning strategy when playing second.*

(3) *If the shadow contains a nonzero number of NSIs and an odd number of SIs, then the linker has a winning strategy when playing first.*

Proof. We begin by proving statement (1). Since the two-component link shadow contains no NSIs, the game starts on a split two-component link shadow and the unlinker wins automatically because resolving crossings will never create an unsplittable link diagram.

We now prove statement (2). Assign an arbitrary orientation to the two-component link shadow. If the unlinker plays on an SI, the linker should respond by also playing

on an SI. Since the number of SIs is even, the linker will always be able to respond on a remaining SI until all SIs are resolved. Note that the SIs do not affect the (pseudo)linking number, so how the linker responds is arbitrary.

Recall that the pseudolinking number of a two-component link shadow (or a two-component link pseudodiagram with only SIs resolved) is 0. Thus, after the unlinker plays for the first time on an NSI, the two-component link pseudodiagram will have a positive pseudolinking number of $\frac{1}{2}$ or a negative pseudolinking number of $-\frac{1}{2}$. The linker should then respond on any other NSI by resolving the crossing to have the same sign as the crossing resolved by the unlinker on the previous move, which will change the pseudolinking number to 1 or -1. For the remaining NSIs, when the unlinker plays on an NSI, the linker should respond on a remaining NSI by resolving the crossing to have sign opposite to the crossing resolved by the unlinker on the previous move. Since Proposition 2.13 guarantees that every two-component link pseudodiagram contains an even number of NSIs, the linker will always be able to respond on a remaining NSI until all NSIs are resolved.

Notice that the linker's response strategy preserves the pseudolinking number at 1 or -1 (after their first response on an NSI) for the remainder of the game. Consequently, when the linker makes the final move, the resulting two-component link diagram will have a nonzero linking number. This implies that the link diagram is unsplittable and, therefore, the linker wins.

We now prove statement (3). Assign an arbitrary orientation to the two-component link shadow. The linker should begin by playing arbitrarily on an SI. This will effectively reduce the game to one with an even number of unresolved SIs where the linker is now playing second on this new link pseudodiagram. But then the linker can use the strategy from the proof of statement (2) to guarantee a win. □

Conclusion

In Section 3 of this paper, we introduced a new two-player combinatorial game played on shadows of two-component link diagrams, called the linking-unlinking game, where players take turns resolving crossings. One player, the linker, wins if the resulting two-component link diagram is unsplittable, while the other player, the unlinker, wins if the resulting two-component link diagram is splittable.

By studying the decomposition of a rational tangle word into SIs and NSIs, by applying the R2- and anti-R2-strategies, and by utilizing tangle equivalences, we were able to find winning strategies for playing the linking-unlinking game on any shadow of a two-component rational tangle closure (see Theorems 3.9(1) and 3.12).

By studying the pseudolinking numbers of general two-component link shadows, we were able to find winning strategies for the unlinker in the pathological case when the shadow contains no NSIs and winning strategies for the linker in half of

the cases when the shadow contains NSIs (see Theorem 3.14). Recall that ensuring a linking number of 0 does not guarantee that a link diagram will be splittable. This suggests that linking number arguments may not be as useful when seeking winning strategies for the unlinker in the remaining half of the cases when the shadow contains NSIs. As such, these cases currently remain open.

Question 3.15. Can explicit winning strategies for playing the linking-unlinking game on two-component link shadows be found in the remaining cases not covered by Theorem 3.14?

If winning strategies for playing the linking-unlinking game on all two-component link shadows cannot be found, then it would be interesting to find families of two-component link shadows beyond rational tangle closures for which winning strategies can be found.

Question 3.16. For which new families of link shadows can explicit winning strategies for playing the linking-unlinking game be found?

References

[Adams 1994] C. C. Adams, *The knot book: an elementary introduction to the mathematical theory of knots*, Freeman, New York, 1994. MR Zbl

[Bollobás 1998] B. Bollobás, *Modern graph theory*, Graduate Texts in Math. **184**, Springer, 1998. MR

[Brown et al. 2017] S. Brown, F. Cabrera, R. Evans, G. Gibbs, A. Henrich, and J. Kreinbihl, "The region unknotting game", *Math. Mag.* **90**:5 (2017), 323–337. MR Zbl

[Ganzell et al. 2014] S. Ganzell, A. Meadows, and J. Ross, "Twist untangle and related knot games", *Integers* **14** (2014), art. id. G4. MR Zbl

[Hanaki 2010] R. Hanaki, "Pseudo diagrams of knots, links and spatial graphs", *Osaka J. Math.* **47**:3 (2010), 863–883. MR Zbl

[Henle 1979] M. Henle, *A combinatorial introduction to topology*, Freeman, San Francisco, 1979. MR Zbl

[Henrich et al. 2011] A. Henrich, N. MacNaughton, S. Narayan, O. Pechenik, R. Silversmith, and J. Townsend, "A midsummer knot's dream", *College Math. J.* **42**:2 (2011), 126–134. MR Zbl

[Johnson 2011] W. Johnson, "The knotting-unknotting game played on sums of rational shadows", preprint, 2011. arXiv

[Johnson and Henrich 2017] I. Johnson and A. K. Henrich, *An interactive introduction to knot theory*, Dover, Mineola, NY, 2017. Zbl

Received: 2018-07-30 Revised: 2019-05-17 Accepted: 2019-06-11

agiambrone@elmira.edu *Elmira College, Elmira, NY, United States*

jmurp61@lsu.edu *Louisiana State University, Baton Rouge, LA, United States*

INVOLVE 12:7 (2019)
dx.doi.org/10.2140/involve.2019.12.1143

On generalizing happy numbers to fractional-base number systems

Enrique Treviño and Mikita Zhylinski

(Communicated by Kenneth S. Berenhaut)

Let n be a positive integer and $S_2(n)$ be the sum of the squares of its decimal digits. When there exists a positive integer k such that the k-th iterate of S_2 on n is 1, i.e., $S_2^k(n) = 1$, then n is called a happy number. The notion of happy numbers has been generalized to different bases, different powers and even negative bases. In this article we consider generalizations to fractional number bases. Let $S_{e,p/q}(n)$ be the sum of the e-th powers of the digits of n base $\frac{p}{q}$. Let k be the smallest nonnegative integer for which there exists a positive integer $m > k$ satisfying $S^k_{e,p/q}(n) = S^m_{e,p/q}(n)$. We prove that such a k, called the height of n, exists for all n, and that, if $q = 2$ or $e = 1$, then k can be arbitrarily large.

1. Introduction

Let n be a positive integer and $S_2(n)$ be the sum of the squares of its decimal digits. It is well known (for a complete proof look at [Honsberger 1970]) that if you apply a sufficiently high iterate of S_2 to n, the result is either 1 or is in the cycle

$$4 \to 16 \to 37 \to 58 \to 89 \to 145 \to 42 \to 20 \to 4.$$

If the iteration reaches 1, we say n is *happy*. A natural generalization is to allow for any base-b representation of the digits, where $b \geq 2$, and to replace the sum of squares of digits with the sum of e-th powers of the digits for some integer $e \geq 1$. Let $S_{e,b}(n)$ be the sum of e-th powers of the digits of n when n is written in base b. If there exists an integer k such that $S^k_{e,b}(n) = 1$, we say n is an e-power b-happy number (when $e = 2$, we call n a b-happy number). Suppose that there exist integers k and m with $0 \leq k < m$ such that $S^k_{e,b}(n) = S^m_{e,b}(n)$; then the iterates of n under $S_{e,b}$ will cycle through the sequence $\{S^k_{e,b}(n), S^{k+1}_{e,b}(n), \ldots, S^{m-1}_{e,b}(n)\}$. If $m-k$ is minimal, then we say that n reaches the cycle $(S^k_{e,b}(n), S^{k+1}_{e,b}(n), \ldots, S^{m-1}_{e,b}(n))$. If k is the smallest nonnegative integer for which this is true, we say k is the *height* of n.

MSC2010: 11A63.
Keywords: happy numbers, fractional base, digital representation.

The study of which cycles can be reached for $e \in \{2, 3\}$ and $2 \le b \le 10$ was done in [Grundman and Teeple 2001]. The techniques in that paper can easily be used to find the cycles for other choices of e and b. Another generalization is to allow the base b to be a negative number. It turns out that for a positive integer n, there is a unique set of digits $0 \le a_i \le |b| - 1$ such that $n = \sum_{i=0}^{r} a_i b^i$. Grundman and Harris [2018] found the cycles reached for $-2 \ge b \ge -10$ and $e = 2$. The authors also study in what cases there exist consecutive b-happy numbers in an arithmetic progression, generalizing [El-Sedy and Siksek 2000; Grundman and Teeple 2007].

Bland et al. [2017] addressed a series of questions regarding a generalization of happy numbers to fractional bases. For integers $p > q > 0$ with $\gcd(p, q) = 1$, each positive integer n has a unique representation in base $\frac{p}{q}$. Namely, there exists an integer $r \ge 0$ such that for every integer $i \in \{0, 1, \ldots, r\}$ there exists an integer $a_i \in \{0, 1, \ldots, p-1\}$ with $a_r \ne 0$ and

$$n = \sum_{i=0}^{r} a_i \left(\frac{p}{q}\right)^i.$$

For our notation, we will say $n = \overline{a_r a_{r-1} \cdots a_1 a_0}_{p/q}$. Let $S_{e,p/q}(n)$ be the sum of the e-th powers of the digits of n when written in fractional base $\frac{p}{q}$; i.e.,

$$S_{e,p/q}(n) = \sum_{i=0}^{r} a_i^e.$$

In [Bland et al. 2017], the authors studied the case when $e = 2$ and proved that there are no happy numbers greater than 1 for any fractional base. They mainly study the fractional base $\frac{3}{2}$, finding the possible cycles that $S_{2,3/2}$ can reach. They end the paper with several questions. The three we will focus on in this paper are the following:

(1) Can we find the cycles reached by $S_{e,b}$ for different e-th powers when $\frac{p}{q} = \frac{3}{2}$?

(2) Can we find the cycles reached by $S_{e,b}$ for different $\frac{p}{q}$ when we restrict to $e = 2$?

(3) Are there positive integers n of arbitrarily large height?

In the case of positive integer bases, that there are numbers with arbitrary height is relatively simple to prove because you can find an n such that $S_{e,b}(n) = k$ for any positive integer k by having n be a number with k 1's in its base-b expansion. For example, let $a_1 = 10$; then a_1 has height 1 since $S_2(10) = 1$. Let

$$a_2 = \underbrace{11\ldots1}_{10}.$$

Since $S_2(a_2) = 10$, a_2 has height 2. Let

$$a_n = \underbrace{11\ldots1}_{a_{n-1}}.$$

Then a_n has height n. This simple process creates a sequence of numbers with larger and larger heights by attaching the appropriate number of 1's to a number. The problem with fractional bases is that not every choice of digits leads to an integer. For example $\overline{11}_{3/2}$ is not an integer, since $1 + \frac{3}{2} \notin \mathbb{Z}$.

We answer the three questions with two theorems. The first theorem answers two of the questions.

Theorem 1. *Let $p > q$ be positive integers with $\gcd(p, q) = 1$, and let e be a positive integer. Then, for every positive integer n, the repeated iterations of the function $S_{e,p/q}$ on n will eventually reach a cycle. In particular, the possible cycles reached for $1 \le e \le 12$, $\frac{p}{q} = \frac{3}{2}$ can be found in Table 2, answering the first question. Also, the possible cycles reached for $e \in \{2, 3, 4\}$ and $\frac{p}{q} \in \left\{\frac{5}{2}, \frac{5}{3}, \frac{5}{4}, \frac{7}{2}\right\}$ are in Table 3, answering the second question.*

The second theorem answers the third question for a special class of fractional bases that includes $\frac{3}{2}$, and for all fractional bases when $e = 1$.

Theorem 2. *Let $p > q$ be positive integers with $\gcd(p, q) = 1$, and let e and H be positive integers. If $q = 2$ or $e = 1$, then there exists an integer n such that the height of n is H.*

In Section 2, we will present useful background on fractional-base number systems. In Section 3, we prove Theorem 1. Finally, in Section 4, we prove Theorem 2.

2. Fractional-base number systems

As mentioned in the Introduction, for any $\frac{p}{q}$ with $\gcd(p, q) = 1$ and $p > q$, for every positive integer n, there exist *fractional digits* $a_0, a_1, \ldots, a_r$ satisfying $0 \le a_i < p$ for $i \in \{0, 1, \ldots, r-1\}$ and $0 < a_r < p$ such that

$$n = \sum_{i=0}^{r} a_i \left(\frac{p}{q}\right)^i.$$

We will use the following notation to denote that a_i are the fractional digits of n base $\frac{p}{q}$:

$$n = \overline{a_r a_{r-1} a_{r-2} \ldots a_2 a_1 a_0}_{p/q}.$$

For example base $\frac{3}{2}$ uses numbers 0, 1, 2 as digits. Table 1 gives the base-$\frac{3}{2}$ representations of some decimal numbers.

It is easy to find n given its expansion in base $\frac{p}{q}$, but going the other way around is a little harder. Suppose we have the number n and we want to find its fractional digits base $\frac{p}{q}$. Let $n = \overline{a_r a_{r-1} \ldots a_1 a_0}_{p/q}$. Then

$$n - a_0 = \left(\frac{p}{q}\right) \overline{a_r a_{r-1} \ldots a_1}_{p/q}.$$

n	n in base $\frac{3}{2}$	n	n in base $\frac{3}{2}$
0	$\bar{0}_{3/2}$	6	$\overline{210}_{3/2}$
1	$\bar{1}_{3/2}$	7	$\overline{211}_{3/2}$
2	$\bar{2}_{3/2}$	8	$\overline{212}_{3/2}$
3	$\overline{20}_{3/2}$	9	$\overline{2100}_{3/2}$
4	$\overline{21}_{3/2}$	10	$\overline{2101}_{3/2}$
5	$\overline{22}_{3/2}$	11	$\overline{2102}_{3/2}$

Table 1. The first 12 nonnegative integers in the $\frac{3}{2}$-base number system.

The left side is an integer, so the right side is also an integer. Since $\gcd(p, q) = 1$, $q \mid \overline{a_r a_{r-1} \dots a_1}_{p/q}$, and so $p \mid (n - a_0)$. Therefore $n \equiv a_0 \mod p$. There is a unique a_0 in $\{0, 1, 2, \dots, p-1\}$ that is congruent to n modulo p. But we also have

$$\overline{a_r a_{r-1} \dots a_1}_{p/q} = \left(\frac{q}{p}\right)(n - a_0).$$

We repeat the process and we can say that

$$n \equiv \left(\frac{q}{p}\right)(n - a_0) - a_1 \mod p.$$

Therefore, we can find a_1. We can repeat this process until we reach 0 and find all of the digits of n.

We can summarize the algorithm to translate numbers into the fractional base $\frac{p}{q}$ as follows:

(1) Compute $n_0 = n \pmod p$.

(2) Compute $n = (n - n_0)\left(\frac{q}{p}\right)$.

(3) Repeat steps 1 and 2, until n is zero.

As an example, suppose we want to find the digits of 12 in base $\frac{3}{2}$. First we have $12 \equiv 0 \mod 3$, so $a_0 = 0$. Then we calculate $(12 - 0)\frac{2}{3} = 8$. We find $8 \equiv 2 \mod 3$, so $a_1 = 2$. Then we find $(8 - 2)\frac{2}{3} = 4$ and $4 \equiv 1 \mod 3$, so $a_2 = 1$. Then we find $(4 - 1)\left(\frac{2}{3}\right) = 2$ and $2 \equiv 2 \mod 3$, so $a_3 = 2$. Since the next step yields 0, we've found that $12 = \overline{2120}_{3/2}$.

3. The cycles formed when iterating $S_{e,3/2}$

An integer $n > 1$ cannot be *happy* in a fractional-base number system. Indeed suppose that n is e-power $\frac{p}{q}$-happy; then $S^m_{e,p/q}(n) = 1$ for some minimal positive integer m. But then $k = S^{m-1}_{e,p/q}(n)$ must satisfy that the sum of the e-th powers of its digits is 1. Therefore the fractional-base expansion of k is $\overline{100\dots0}_{p/q}$. But that

e	cycles	n^*
1	(1), (2)	2
2	(1), (5, 8, 9)	8
3	(1), (9), (10), (17, 18)	32
4	(1), (51), (52)	77
5	(1), (131), (98, 99)	185
6	(1), (197, 260, 387, 323, 263, 450), (324, 131, 259)	419
7	(1), (771, 516, 643, 518)	1211
8	(1), (1539, 775, 1284), (1287, 1794, 1796, 2052), (1032), (1033)	2723
9	(1), (2566), (2565)	6557
10	(1), (10247)	13118
11	(1), (14342, 16388, 14344), (14341), (14340)	27968
12	(1), (28678), (28677)	62933

Table 2. Cycles reached when iterating $S_{e,3/2}$, and the value of n^* for different values of e.

means $k = \left(\frac{p}{q}\right)^r$ for some integer r. This number is not an integer unless $r = 0$, which would imply $k = 1$, but we assumed $k > 1$. While happiness is impossible, we can still search which cycles can be reached. For us to be able to prove that the determination of cycles is complete, we need to first prove the following lemma.

Lemma 1. *Let $\frac{p}{q}$ satisfy $p > q$ and $\gcd(p, q) = 1$, and let e be a positive integer. Then, there exists an n^* such that $S_{e,p/q}(n^*) \geq n^*$, and $S_{e,p/q}(n) < n$ for all $n > n^*$.*

The values of n^* for different values of e and $\frac{p}{q} = \frac{3}{2}$ can be found in the last column of Table 2. The values of n^* for $e \in \{2, 3, 4\}$ and $\frac{p}{q} \in \left\{\frac{5}{2}, \frac{5}{3}, \frac{5}{4}, \frac{7}{2}\right\}$ are in Table 3.

Proof. Let n be a positive integer. Then

$$n = \overline{a_r a_{r-1} \dots a_1 a_0}_{p/q} = \sum_{i=0}^{r} a_i \left(\frac{p}{q}\right)^i \geq a_r \left(\frac{p}{q}\right)^r \geq \left(\frac{p}{q}\right)^r,$$

so $r \leq \log_{p/q}(n)$. But then

$$S_{e,p/q}(n) = \sum_{i=0}^{r} a_i^e < \sum_{i=0}^{r} p^e = (r+1)p^e \leq (\log_{p/q}(n) + 1)p^e.$$

Since p^e is a constant, for a large enough n

$$n > (\log_{p/q}(n) + 1)p^e > S_{e,p/q}(n). \tag{1}$$

Indeed, one can confirm with L'Hôpital's rule that $n/(C \log(n)) \to \infty$ as $n \to \infty$ for any constant $C > 0$.

$\frac{p}{q}$	$e=2$	$e=3$	$e=4$
$\frac{5}{2}$	(16, 6, 5, 4), (32, 24, 29) $n^*=39$	(65), (163, 190, 73, 118, 64), (81), (80), (66), (17) $n^*=239$	(371, 276, 275, 274), (355, 130, 113), (195, 353) $n^*=1039$
$\frac{5}{3}$	(34, 50), (25), (26), (59), (23), (11), (10) $n^*=59$	(100, 38, 64, 102, 46), (101, 39), (127, 107, 73, 135), (162), (193), (190, 166, 218), (199, 237) $n^*=284$	(772, 804, 454, 788, 950, 658, 934, 1126, 1028, 1202, 868, 936, 390), (1027, 1137, 1125), (1122, 994), (1299), (101), (100) $n^*=1324$
$\frac{5}{4}$	(66, 55), (50), (58, 75, 49, 56, 67), (74, 83), (51) $n^*=74$	(311, 251, 247, 231, 371), (361), (417), (374), (360), (314), (424, 418, 436, 272, 328, 364) $n^*=464$	(1786, 1880, 1403, 1594, 1659, 2011, 2075, 1579, 2057, 1947, 1688, 1229, 1641, 1676, 1946, 1673, 1851, 1592, 1419, 1974, 2058, 2012, 2090) $n^*=2639$
$\frac{7}{2}$	(25, 52), (97) $n^*=97$	(341, 591, 376, 143, 187, 216, 352, 25, 280, 244, 469, 63, 128, 44, 141, 161, 197, 73, 307, 467, 377, 234, 182, 91), (35), (288, 343, 9, 16, 72), (36), (189), (190), (468) $n^*=615$	(914, 2065, 1953, 1538, 2819, 2690, 2210, 1507, 1491, 2610, 1856, 1348, 1666, 259, 1808, 2659, 3136, 1824), (1634, 1731, 994), (371, 34, 1313), (130, 354, 289, 1938, 3265, 2930, 1474, 1570), (451, 195, 2177, 1554, 179, 513, 2034, 2530) $n^*=5417$

Table 3. Cycles reached when iterating $S_{e,p/q}$, and the value of n^* for different values of e and $\frac{p}{q}$.

Therefore, there is a maximum n^* such that $n^* < S_{e,p/q}(n^*)$.

To calculate the precise value of n^*, we use a computer to find an N for which (1) is satisfied. Then we evaluate $S_{e,p/q}(n)$ for all $n \le N$ and record which one is the largest satisfying $n \le S_{e,p/q}(n)$. □

Proof of Theorem 1. To simplify notation, let $S(n) = S_{e,p/q}(n)$ for all positive integers n. Let n^* be as in Lemma 1. Now, for each $m \leq n^*$, compute $m, S(m), S(S(m)), \ldots$ until it cycles. The process terminates because $S(n) < n$ for all $n > n^*$. Therefore, for $n > n^*$, there exists a positive integer k such that $S^k(n) \leq n^*$. This implies that the cycle n reaches is one that was already computed. Therefore, we need only find the cycles reached for $m \leq n^*$. The outcome of performing these calculations for different values of e and $\frac{p}{q} = \frac{3}{2}$ is recorded in Table 2. The outcome of performing these calculations on $e \in \{2, 3, 4\}$ with $\frac{p}{q} \in \left\{\frac{5}{2}, \frac{5}{3}, \frac{5}{4}, \frac{7}{2}\right\}$ is recorded in Table 3. □

4. Arbitrary heights in fractional-base number systems

The key to our proof of Theorem 2 is showing that for each sufficiently large k, there exists a positive integer n such that $S_{e,p/q}(n) = k$. The following lemma handles the case when $q = 2$.

Lemma 2. *Let $e \geq 1$ and $p > 2$ be an odd positive integer. For every integer $k \geq 2^e$, there exists an even integer n such that $S_{e,p/2}(n) = k$.*

Proof. We will prove the lemma by induction on k. To show that it is true for $k = 2^e$, consider the number 2. Since 2 is $\bar{2}_{p/2}$, we have $S_{e,p/2}(2) = 2^e$. Now let $k \geq 2^e$ and assume that there exists an even m such that $S_{e,p/2}(m) = k$. Let $m = 2^b c$, where $b \geq 1$ and c is odd. Write m in base $\frac{p}{2}$ as

$$m = \overline{a_r a_{r-1} \ldots a_1 a_0}.$$

Then

$$\left(\frac{p}{2}\right)^b m + 1 = \overline{a_r a_{r-1} \ldots a_1 a_0 \underbrace{0 \ldots 0}_{b-1} 1},$$

where there are $b - 1$ zero digits. Since $m = 2^b c$, we know $\left(\frac{p}{2}\right)^b m + 1$ is even. Furthermore, since it has the same digits as before with $b - 1$ zeroes added and one 1 added, the sum of the e-th powers of the digits is $k + 1$. □

The following lemma handles the $e = 1$ case.

Lemma 3. *Let $\frac{p}{q} > 1$ be written in lowest terms. For every integer $k \geq q$, there exists n such that $S_{1,p/q}(n) = k$.*

Proof. We prove by induction on t that for each $k \in \{q, q + 1, \ldots, qt\}$, there exists an m_k such that $S_{1,p/q}(m_k) = k$. The fact that $S_{1,p/q}(q) = q$ proves the case of $t = 1$. Now, fix $t \geq 1$ and assume that for each $k \in \{q, q + 1, \ldots, qt\}$, there exists an m_k such that $S_{1,p/q}(m_k) = k$. Write m_{qt} as $m_{qt} = q^\alpha b$ for some $\alpha \geq 1$ and b

relatively prime to q. Suppose $m_{qt} = \overline{a_r a_{r-1} \dots a_0}_{p/q}$. Then

$$\ell = \left(\frac{p}{q}\right)^{\alpha} m_{qt} = \overline{a_r \dots a_0 \underbrace{0 \dots 0}_{\alpha}}.$$

We know $\ell \not\equiv 0 \bmod q$. Let w be the smallest positive integer such that $\ell + w \equiv 0 \bmod q$. Then $1 \le w \le q-1 < p-1$. But then

$$\ell + w = \overline{a_r \dots a_0 \underbrace{0 \dots 0}_{\alpha-1} w},$$

because $w < p-1$. This implies that the digital sums base $\frac{p}{q}$ of the numbers $\ell+1, \ell+2, \dots, \ell+w$ are $qt+1, qt+2, \dots, qt+w$, respectively. Now $\ell+w$ is a multiple of q with $S_{1,p/q}(\ell+w) = qt+w \ge qt+1$, and we have that for all $q \le k \le qt+w$ there exists m_k such that $S_{1,p/q}(m_k) = k$. Since $q \mid (\ell+w)$ and $\ell+w \ge qt+1$, we have $\ell+w \ge q(t+1)$. Therefore, we've proved that for every $q \le k \le q(t+1)$, there is an m_k such that $S_{1,p/q}(m_k) = k$. □

Using these two lemmas, we can now present the proof of Theorem 2.

Proof of Theorem 2. We will prove it by induction. Let n^* be as defined in Lemma 1. Since the cycles that are reached by iterations of $S_{e,p/q}$ are finite and there are finitely many of them, there is a largest integer K with height 0. Let n be an integer greater than $M = \max\{n^*, 2^e, q, K\}$. Since $n > K$, we know n has some height $h > 0$. Then, $S_{e,p/q}(n)$ has height $h-1$, $S^2_{e,p/q}(n)$ has height $h-2, \dots$, and $S^{h-1}_{e,p/q}(n)$ has height 1. Therefore, for every positive integer $i \le h$, there exists an integer n of height i.

Let $H \ge h$. Suppose that there is an integer $n > K$ with height H. Since $n > 2^e$, by Lemma 2, if $q = 2$, then there exists t such that $S_{e,p/2}(t) = n$. Since $n > q$, by Lemma 3, if $e = 1$, then there exists t such that $S_{1,p/q}(t) = n$. Therefore, in either case ($q = 2$ or $e = 1$), there exists an integer t such that $S_{e,p/q}(t) = n$. But $t > n^*$, which implies that $n = S_{e,p/q}(t) < t$. Therefore $t > n > K$ and t has height $H+1$. □

Acknowledgements

We would like to thank the anonymous referee for making several suggestions which improved the clarity of the paper. We would also like to thank Lake Forest College for funding both authors.

References

[Bland et al. 2017] A. Bland, Z. Cramer, P. de Castro, D. Domini, T. Edgar, D. Johnson, S. Klee, J. Koblitz, and R. Sundaresan, "Happiness is integral but not rational", *Math Horiz.* **25**:1 (2017), 8–11. MR

[El-Sedy and Siksek 2000] E. El-Sedy and S. Siksek, "On happy numbers", *Rocky Mountain J. Math.* **30**:2 (2000), 565–570. MR Zbl

[Grundman and Harris 2018] H. G. Grundman and P. E. Harris, "Sequences of consecutive happy numbers in negative bases", *Fibonacci Quart.* **56**:3 (2018), 221–228. MR Zbl

[Grundman and Teeple 2001] H. G. Grundman and E. A. Teeple, "Generalized happy numbers", *Fibonacci Quart.* **39**:5 (2001), 462–466. MR Zbl

[Grundman and Teeple 2007] H. G. Grundman and E. A. Teeple, "Sequences of consecutive happy numbers", *Rocky Mountain J. Math.* **37**:6 (2007), 1905–1916. MR Zbl

[Honsberger 1970] R. Honsberger, *Ingenuity in mathematics*, New Mathematical Library **23**, Random House, New York, 1970. MR Zbl

Received: 2018-09-30 Revised: 2019-06-12 Accepted: 2019-06-22

trevino@lakeforest.edu *Lake Forest College, Lake Forest, IL, United States*

zhylinskim@mx.lakeforest.edu *Lake Forest College, Lake Forest, IL, United States*

msp

INVOLVE 12:7 (2019)
dx.doi.org/10.2140/involve.2019.12.1153

On the Hadwiger number of Kneser graphs and their random subgraphs

Arran Hamm and Kristen Melton

(Communicated by Anant Godbole)

For $n, k \in \mathbb{N}$, let $\mathrm{KG}(n, k)$ be the usual Kneser graph (whose vertices are k-sets of $\{1, 2, \ldots, n\}$ with $A \sim B$ if and only if $A \cap B = \varnothing$). The Hadwiger number of a graph G, denoted by $h(G)$, is $\max\{t : K_t \preccurlyeq G\}$, where $H \preccurlyeq G$ if H is a minor of G. Previously, lower bounds have been given on the Hadwiger number of a graph in terms of its average degree. In this paper we give lower bounds on $h(\mathrm{KG}(n, k))$ and $h(\mathrm{KG}(n, k)_p)$, where $\mathrm{KG}(n, k)_p$ is the binomial random subgraph of $\mathrm{KG}(n, k)$ with edge probability p. Each of these bounds is larger than previous bounds under certain conditions on k and p.

1. Introduction

Over the past few decades graph parameters of Kneser graphs have been studied extensively. The Kneser graph with parameters n and k has the k-sets of $\{1, 2, \ldots, n\}$ as its vertex set, with $A \sim B$ if and only if $A \cap B = \varnothing$. In particular, the independence number, chromatic number, diameter, and bandwidth parameters have been examined for members of this family (see [Erdős et al. 1961; Lovász 1978; Valencia-Pabon and Vera 2005; Jiang et al. 2017], respectively). In the present paper we continue the study of parameters of Kneser graphs by giving lower bounds on the *Hadwiger number* of Kneser graphs and random subgraphs of Kneser graphs. The Hadwiger number of a graph G, denoted by $h(G)$, is $\max\{t : K_t \preccurlyeq G\}$, where we say $H \preccurlyeq G$ if H is a minor of G and K_t is the complete graph, or *clique*, on t vertices.

To introduce the Hadwiger number, it's worth mentioning one of the most important open problems in graph theory — Hadwiger's conjecture. The conjecture is that if a graph has chromatic number t, then it contains K_t as a minor. It has been shown that this conjecture holds for $t \leq 6$; see [Seymour 2016] for a survey of the problem. A few decades ago, the following were proven which relate the Hadwiger number of a graph to its average degree.

MSC2010: 05C83, 05C80, 05D40.
Keywords: Kneser graphs, Hadwiger number.

Theorem 1.1 [Kostochka 1984]. *There is a constant $c > 0$ such that if G is a graph with average degree $d \geq 2$, then*

$$h(G) \geq c\frac{d}{\sqrt{\ln(d)}}. \tag{1}$$

Theorem 1.2 [Kostochka 1982]. *There is a constant $C > 0$ such that if $\mathcal{G}$ is the set of graphs with average degree d for d sufficiently large, then*

$$\min_{G\in\mathcal{G}} h(G) \leq C\frac{d}{\sqrt{\ln(d)}}. \tag{2}$$

Notice in particular that (1) gives a lower bound on $h(G)$ for graphs with average degree d and (2) implies that up to a constant factor (1) cannot be improved when considering the collection of all graphs with average degree d as long as d is big enough. In this paper we begin by focusing on Kneser graphs $\left(\text{for which } d = \binom{n-k}{k}\right)$ and prove the following theorem in Section 2.

Theorem 1.3. *Suppose $n = t(k^2+k)+r$ for natural numbers t and r with $0 \leq r \leq k^2+k-1$. Then*

$$h(\mathrm{KG}(n,k)) \geq \frac{1}{k+1}\binom{n-r}{k}. \tag{3}$$

In particular, when k is small compared to n (up to about $\ln(n)$) the lower bound in (3) exceeds that in (1). More precisely, suppose $k \ll \ln(n)$ (where we say $f(n) \ll g(n)$ or, equivalently, $f(n) = o(g(n))$ if $\lim_{n\to\infty} f(n)/g(n) = 0$). Then for $\mathrm{KG}(n,k)$ the bounds in Theorems 1.1 and 1.3 are, up to a constant factor, $\binom{n}{k}/\sqrt{k\ln(n)}$ and $\binom{n}{k}/k$, respectively. So in this case the coefficient of $\binom{n}{k}$ in Theorem 1.3 is larger, which verifies the claim.

We next consider the Hadwiger number for binomial random subgraphs of Kneser graphs. Since the introduction of the Erdős–Rényi random graph [1959], there has been interest in finding parameters of binomial random graphs. For context, the Erdős–Rényi random graph results from forming a binomial random subgraph of the complete graph; over the past couple of decades binomial random subgraphs underlying other graphs, specifically Kneser graphs, have been examined. In particular, the independence number, see [Bollobás et al. 2016; Devlin and Kahn 2016], and the chromatic number, see [Kupavskii 2016], for this type of random graph have been studied. We further this study by obtaining a lower bound on $h(\mathrm{KG}(n,k)_p)$ (where $\mathrm{KG}(n,k)_p$ is the binomial random subgraph of $\mathrm{KG}(n,k)$ with edge probability p) in the following theorem.

Theorem 1.4. *Let $k \ll \sqrt{n}$ and $N = \binom{n}{k}$. If m and p satisfy $\sqrt{\ln(N)} \ll m \ll n/k$, $2k \leq m$, and $p \gg \max\{\ln(m)/m, \ln(N)/m^2\}$, then*

$$h(\mathrm{KG}(n,k)_p) \geq \frac{1}{2m}\binom{n}{k} \text{ w.h.p.} \tag{4}$$

As is standard, for an event E depending on the (often hidden) parameter n, we say E occurs *with high probability* (*w.h.p.*) if $\Pr(E) \to 1$ as $n \to \infty$.

Additionally, we obtain the following corollary which relates the bound of Theorem 1.4 to (1).

Corollary 1.5. *For each $k \ll \sqrt{n}$, there are values of m and p (as in Theorem 1.4) such that the lower bound on $h(\mathrm{KG}(n,k)_p)$ in (4) exceeds that of (1).*

We will prove Theorem 1.4 and Corollary 1.5 in Section 4 after giving some preliminary notation and lemmas in Section 3. Finally, in Section 5 we state a generalization of Theorem 1.4 and mention a couple of open problems.

2. Proof of Theorem 1.3

Before presenting the proof of Theorem 1.3, we need a couple of preliminaries. First, as is standard, we will let $[n] := \{1, \dots, n\}$. We will also need the following theorem, sometimes referred to as Baranyai's theorem, which gives the existence of a particular decomposition of the collection of k-sets of $[n]$.

Theorem 2.1 [Baranyai 1975]. *If $k \mid n$, there are perfect matchings $\mathcal{A}_i$ such that*

$$\binom{[n]}{k} = \bigsqcup_{i=1}^{\binom{n-1}{k-1}} \mathcal{A}_i.$$

By its very statement, Baranyai's theorem concerns *set systems*. In this context, a *perfect matching* is a collection of k-sets of $[n]$ which are pairwise disjoint and whose union is $[n]$. The conclusion of Baranyai's theorem, then, is that we may partition the collection of all k-sets of $[n]$ into perfect matchings. Of course, we can view the collection of k-sets of $[n]$ as vertices of the Kneser graph with parameters n and k. With this perspective a perfect matching from Baranyai's theorem is a clique on n/k vertices in $\mathrm{KG}(n,k)$. As such, if we have that $k \mid n$, then we can use Theorem 2.1 to give a partition of $V(\mathrm{KG}(n,k))$ into complete graphs each on n/k vertices.

Proof of Theorem 1.3. Let n, k, and r be natural numbers such that $n = t(k^2+k)+r$, where $0 \le r \le k^2+k-1$, and take $G' = \mathrm{KG}(n-r,k)$. Notice that $G' \le G$ and so any minor of G' is a minor of G. Since $k \mid n-r$, Theorem 2.1 applies to G' yielding $K_1, K_2, \dots, K_T$, where each K_i is a complete graph on $(n-r)/k$ vertices and $T = \binom{n-r-1}{k-1}$. By the divisibility assumption, we can partition the vertices of each K_i (arbitrarily) into sets of size $k+1$ which yields $(n-r)/(k(k+1))$ "clusters" for each i (and thus $(n-r)T/(k(k+1))$ total clusters). Note that each cluster is a clique of size $k+1$ in G'. This gives that every vertex within a cluster has at least one edge to every other cluster. This is so because if v is a vertex of G' and W is cluster not containing v, then v can have nonempty intersection with at most k vertices of W (since the $k+1$ vertices in W are pairwise disjoint) and therefore v

is disjoint from at least one of the vertices in W. So if we contract each cluster to a vertex, the result is a complete graph on $(n-r)T/(k(k+1)) = (1/(k+1))\binom{n-r}{k}$ vertices, which is the desired lower bound. □

3. Preliminaries for Theorem 1.4

In this section we gather a few preliminaries from the study of random graphs which we will need to prove Theorem 1.4. The first gives the threshold value for p which ensures that the random graph $G_{n,p}$ is connected.

Theorem 3.1 [Erdős and Rényi 1959]. *If $p \gg \log(n)/n$, then $G_{n,p}$ is connected w.h.p.*

We will also need a couple of basic probability inequalities. The first is standard and the second is stated in the form of Theorem 2.1 in [Janson et al. 2000].

Theorem 3.2 (Markov's inequality). *If $\mathbb{X}$ is a nonnegative random variable and $a > 0$, then*

$$\Pr(\mathbb{X} \geq a) \leq \frac{\mathbb{E}[\mathbb{X}]}{a}.$$

Theorem 3.3 (Chernoff bound). *If $\mathbb{X}$ is the sum of n independent indicator random variables and $0 < \delta < 1$, then*

$$\Pr(\mathbb{X} \leq (1-\delta)\mathbb{E}[\mathbb{X}]) \leq \exp\left[-\tfrac{1}{2}\delta^2 \mathbb{E}[\mathbb{X}]\right].$$

Throughout the paper, we use big O notation in the standard way and make repeated use of:

$$\text{if } k \ll \sqrt{n}, \text{ then } \binom{n-k}{k} \sim \binom{n}{k} \text{ as } n \to \infty.$$

For disjoint vertex sets X and Y of a graph G, let $\nabla_G(X, Y)$ be the set of edges of G with one vertex in X and the other in Y; if the underlying graph G is understood, we will neglect the subscript. In the context of taking a binomial random subgraph of G, we will let $\nabla_p(X, Y)$ be the set of edges in G_p with one vertex in X and the other in Y in order to avoid having a double subscript. For the remainder of the paper, we will take $G := \mathrm{KG}(n, k)$. Before proceeding to the proofs of Theorems 1.3 and 1.4, we will need the following lemma.

Lemma 3.4. *If X and Y are disjoint cliques of size m in G, then $|\nabla(X, Y)| \geq m^2 - km$.*

Proof. For each $x \in X$, x has nonempty intersection with at most k vertices in Y (since the vertices of Y form a clique in G). So $d_Y(x) \geq m - k$ and summing over all vertices in X proves the claim. □

4. Proofs for Theorem 1.4 and Corollary 1.5

Before proceeding, we recall the statement and discuss how the proof will unfold.

Theorem 1.4. *Let $k \ll \sqrt{n}$ and $N = \binom{n}{k}$. If m and p satisfy $\sqrt{\ln(N)} \ll m \ll n/k$, $2k \leq m$, and $p \gg \max\{\ln(m)/m, \ln(N)/m^2\}$, then*

$$h(\mathrm{KG}(n,k)_p) \geq \frac{1}{2m}\binom{n}{k} \text{ w.h.p.} \tag{4}$$

Our strategy is to first use Baranyai's theorem to give a clique decomposition of G into somewhat "large" cliques as in the proof of Theorem 1.3. After randomizing the edges of G, the size of these cliques along with the assumed lower bound on p will ensure that *most* of them will be connected in G_p; each which remains connected will be contracted to a vertex. Next we will use Lemma 3.4 and the size of the cliques to say that if p is big enough, then every pair of cliques will have at least one edge between them in G_p. These two observations combine to say that $h(G_p)$ is at least the number of cliques which remain connected after randomizing edges. Then to prove Corollary 1.5, we simply must choose parameters so that the lower bound in (4) is greater than the lower bound in (1).

We now turn to the proof.

Proof of Theorem 1.4. For ease of reading, recall that $k \ll \sqrt{n}$, $N = \binom{n}{k}$, m satisfies $\sqrt{\ln(N)} \ll m \ll n/k$ and $k \leq m/2$, and $p \gg \max\{\ln(m)/m, \ln(N)/m^2\}$. Fix $\varepsilon > 0$. Since $k \ll \sqrt{n}$ implies $N \sim \binom{n-k}{k}$, we may assume that $k \mid n$, otherwise our argument would, as in the proof of Theorem 1.3, pass to $G' \leq G$ so that the divisibility assumption would be met. Now we may apply Theorem 2.1 to obtain cliques $W_1, \ldots, W_T$, where $T = \binom{n-1}{k-1}$ and $|W_i| = n/k$ for each i. Each of these cliques can be (arbitrarily) partitioned into cliques of size m yielding cliques $U_1, \ldots, U_S$, where $S = N/m$. We will now form G_p by sampling edges in two rounds; first we will sample edges within each U_i and thereafter will sample the remaining edges in G. By Theorem 3.1 (note that $m = \omega_n(1)$) and the first lower bound on p, we have that $G_p[U_i]$ (which is the induced subgraph of G_p on U_i) is connected with probability at least $1 - \varepsilon$ provided that n is large enough. Since S is so large, it is unlikely that every $G_p[U_i]$ is connected; instead we will show that

$$\text{at least half of the } G_p[U_i]\text{'s are connected w.h.p.} \tag{5}$$

This gives away quite a bit, but the "loss" only affects our lower bound by a constant factor.

To prove (5), let $X = \sum X_i$, where each X_i is the indicator of the event $\{G_p[U_i] \text{ is connected}\}$. Thus (5) is the same as

$$\Pr\left(X < \frac{N}{2m}\right) \to 0. \tag{6}$$

Using linearity of expectation and the lower bound on $\Pr(X_i = 1)$ mentioned before (5), we have $\mathbb{E}[X] > (1-\varepsilon)N/m$. Note that these events are independent since they depend on disjoint sets of edges. So, using Theorem 3.3, we obtain

$$\Pr\Big(X < \frac{N}{2m}\Big) \le \Pr\big(X \le \big(1-\tfrac{1}{3}\big)\mathbb{E}[X]\big) \le \exp\big[-\big(\tfrac{1}{3}\big)^2 \tfrac{1}{2}\mathbb{E}[X]\big].$$

So as long as m is chosen so that $N/m \to \infty$ as $n \to \infty$ (and hence $\mathbb{E}[X] \to \infty$ as $n \to \infty$), the right-hand side will tend to zero as $n \to \infty$, which gives (6).

It remains to show,

$$\text{for } i \neq j, \quad |\nabla_p(U_i, U_j)| \neq 0 \text{ w.h.p.} \tag{7}$$

To do so, we let $Y = \sum Y_{i,j}$, where each $Y_{i,j}$ is the indicator of the event $\{\nabla_p(U_i, U_j) = \varnothing\}$ for $i \neq j$. Thus (7) is the same as $\Pr(Y > 0) \to 0$. For this, we have

$$\Pr(Y>0) \le \mathbb{E}[Y] \le \binom{N/m}{2}(1-p)^{(m^2-km)} \le \frac{1}{2m^2}N^2 e^{-p(m^2-km)},$$

where the first inequality comes from Theorem 3.2, the second inequality from linearity of expectation and Lemma 3.4, and the third inequality from the fact that $1-p \le e^{-p}$. The right-hand side can be bounded by

$$N^2 e^{-p(m^2-km)} = \exp[2\ln(N) - p(m^2-km)].$$

The right-hand side tends to zero if and only if $p \gg \ln(N)/(m^2-km)$, which we have by assumption so long as $m \gg \sqrt{\ln(N)}$ and $m \ge 2k$. Indeed for such m, we can choose p appropriately so that the conditions of Theorem 1.4 are satisfied.

So to summarize, provided that $p \gg \ln(m)/m$ there are at least S pods that remain connected after the first round of randomization, where

$$S \ge \frac{1}{2m}\binom{n}{k}.$$

We will then contract all pods which are connected to a vertex and delete all vertices in pods which are disconnected. Provided that $p \gg \ln(N)/m^2$, there is an edge between every pair of remaining vertices and so $h(G_p) \ge S$ as desired. □

We conclude this section by proving Corollary 1.5 after recalling its statement.

Corollary 1.5. *For each $k \ll \sqrt{n}$, there are values of m and p (as in Theorem 1.4) such that the lower bound on $h(\mathrm{KG}(n,k)_p)$ in (4) exceeds that of (1).*

Proof of Corollary 1.5. In order to give parameter values so that the lower bound of (4) is greater than that of (1), we will first need that the average degree in G_p is $(1+o(1))\binom{n-k}{k}p$ (which follows from a straightforward application of Theorem 3.3

on the number of edges in G_p). So for G_p, (1) is

$$\frac{c\binom{n-k}{k}p}{\sqrt{\ln\left(\binom{n-k}{k}p\right)}}, \tag{8}$$

where c is some positive constant.

Notice that the bound of (4) is largest when m is as small as possible and the bound of (1) shrinks with p. Since $k \ll \sqrt{n}$, we have $\binom{n-k}{k} \sim N$. Before defining p, we now must consider two cases which depend on k. If there is some $0 < \alpha < \frac{1}{2}$ so that $k = (1+o(1))n^\alpha$, then we will take $m = n^\beta$, where β satisfies $\frac{1}{2}\alpha < \beta < 1-\alpha$; this is a nonempty interval since $0 < \alpha < \frac{1}{2}$ and a straightforward calculation shows that such an m satisfies the assumptions of Theorem 1.4. In this case, the larger bound on p in the conditions of Theorem 1.4 is $\ln(m)/m$. It is routine to check that for $p \gg \ln(m)/m$, $\ln\left(\binom{n-k}{k}p\right)$ is at most a constant multiple of $\ln(N)$ and so (8) is on the order of $Np/\sqrt{\ln(N)}$. This means the bound of (4) exceeds the bound of (1) for G_p provided that $p \ll \sqrt{\ln(N)}/m$. It remains to verify that $\ln(m)/m \ll \sqrt{\ln(N)}/m$ (i.e., a suitable p-value may be designated). For this observe that $k = (1+o(1))n^\alpha$ and so $\ln(N) \sim k\ln(n/k)$, which means $\sqrt{\ln(N)}/m$ is a constant multiple of $\ln(n)/n^{\beta-(\alpha/2)}$. On the other hand, by the choice of m, $\ln(m)/m = O(\ln(n)/n^\beta)$, which gives the desired relation.

If $k = o(n^\alpha)$ for every $0 < \alpha < \frac{1}{2}$, then in order to define p, let $f(n)$ be some slowly growing function which tends to infinity with n and take $m = f(n)\sqrt{\ln(N)}$. For such m, we have $\ln(m)/m \leq \ln(N)/m^2$ and so the restriction on p in Theorem 1.4 is $p \gg \ln(N)/m^2$. Similar to the previous case, we have that for $p \gg \ln(N)/m^2$, $\ln\left(\binom{n-k}{k}p\right)$ is at most a constant multiple of $\ln(N)$. So (4) exceeds the bound of (1) for G_p provided that $p \ll \sqrt{\ln(N)}/m = 1/(f(n))$. Finally notice that for m-values like this, $1/(f(n)) < 1$, which means that $\ln(N)/m^2 = 1/(f(n))^2 \ll 1/(f(n))$ since $f(n) \to \infty$ as $n \to \infty$, and so such p-values may be chosen. □

5. Concluding remarks

We should note that the proof of Theorem 1.4 presented above gives rise to a slightly more general statement which is:

Theorem 5.1. *Suppose $\{G_n\}$ is an infinite family of graphs such that G_n has $g(n)$ vertices, where $g(n) \to \infty$ with n. Suppose there is an $m(n)$ such that $m := m(n) \to \infty$ with n so that*

(1) *the vertex set of G_n can be partitioned into $\{V_i\}_{i=1}^t$ (where $t = n/m$),*

(2) *$G_n[V_i] \cong K_m$ for each $i = 1, \ldots, t$,*

(3) *$t \to \infty$ with n, and*

(4) *$|\nabla(V_i, V_j)| \geq cm^2$ for $i \neq j$ and $c > 0$ a constant independent of n.*

Then if $\varepsilon > 0$ *and* $p \gg \max\{\ln(m)/m, \ln(n)/m^2\}$ *as* $n \to \infty$, *then w.h.p.* $K_S \preccurlyeq G_p$, *where* $S = (1-\varepsilon)t$.

Because the proof of Theorem 5.1 follows the proof of Theorem 1.4 with only the obvious modifications necessary, we will omit it. We remark that the statement applies naturally to the family of complete balanced m-partite graphs (where the size of each part is parametrized to tend to infinity) and to graph products which are fairly dense and admit a clique decomposition (e.g., $(K_n \,\square\, K_n)^C$).

We conclude by mentioning a couple of open questions. First, we will return to the Hadwiger number of Kneser graphs. As remarked above, the conclusion of Theorem 1.3 only exceeds the bound in Theorem 1.1 if $k \ll \ln(n)$. It may, therefore, be worthwhile to examine the other end of the spectrum, namely the case $n = 2k+1$. In this case, (1) gives that $h(\mathrm{KG}(2k+1, k))$ is bounded below by roughly $k/\ln(k)$. This is not, in general, best possible; if $\binom{2k+1}{k}$ is even (e.g., if k is a power of two), then $\mathrm{KG}(2k+1, k)$ has a 1-factor. A straightforward calculation shows that if we contract each edge of any 1-factor, then the resulting graph is $2k$-regular, which shows that the lower bound in (1) can be effectively doubled. This naturally gives rise to the following question.

Question 5.2. *What is the order of magnitude for* $h(\mathrm{KG}(2k+1, k))$?

Second, it is worth pointing out that the proof of Theorem 1.4 requires that $k \ll \sqrt{n}$. This is because if $k \gg \sqrt{n}$, then the cliques given from Baranyai's theorem are of size n/k, which is much smaller than k. For these cliques of size n/k, the conclusion of Lemma 3.4 becomes trivial, meaning that in this case we cannot exploit the property that every pair of cliques has many edges between them. It seems plausible that either choosing the clique decomposition carefully or using another decomposition of $\mathrm{KG}(n, k)$ (e.g., into complete bipartite graphs) may yield a suitable edge count between pieces of the decomposition like the bound Lemma 3.4, but we have not pursued this. We, therefore, put forth the following question.

Question 5.3. *Can* (1) *be improved for* $(\mathrm{KG}(n, k))_p$ *if* $k \gg \sqrt{n}$?

Acknowledgements

This project was supported by an Institutional Development Award (IDeA) from the National Institute of General Medical Sciences (2 P20 GM103499 15) from the National Institutes of Health. Additionally, we would like to thank the referee for thoughtful suggestions.

References

[Baranyai 1975] Z. Baranyai, "On the factorization of the complete uniform hypergraph", pp. 91–108 in *Infinite and finite sets, I* (Keszthely, Hungary, 1973), edited by A. Hajnal et al., Colloq. Math. Soc. János Bolyai **10**, North-Holland, Amsterdam, 1975. MR Zbl

[Bollobás et al. 2016] B. Bollobás, B. P. Narayanan, and A. M. Raigorodskii, "On the stability of the Erdős–Ko–Rado theorem", *J. Combin. Theory Ser. A* **137** (2016), 64–78. MR Zbl

[Devlin and Kahn 2016] P. Devlin and J. Kahn, "On 'stability' in the Erdős–Ko–Rado theorem", *SIAM J. Discrete Math.* **30**:2 (2016), 1283–1289. MR Zbl

[Erdős and Rényi 1959] P. Erdős and A. Rényi, "On random graphs, I", *Publ. Math. Debrecen* **6** (1959), 290–297. MR Zbl

[Erdős et al. 1961] P. Erdős, C. Ko, and R. Rado, "Intersection theorems for systems of finite sets", *Quart. J. Math. Oxford Ser.* (2) **12** (1961), 313–320. MR Zbl

[Janson et al. 2000] S. Janson, T. Łuczak, and A. Ruciński, *Random graphs*, Wiley, New York, 2000. MR Zbl

[Jiang et al. 2017] T. Jiang, Z. Miller, and D. Yager, "On the bandwidth of the Kneser graph", *Discrete Appl. Math.* **227** (2017), 84–94. MR Zbl

[Kostochka 1982] A. V. Kostochka, "On the minimum of the Hadwiger number for graphs with a given mean degree of vertices", *Metody Diskret. Analiz.* **38** (1982), 37–58. In Russian; translated in Amer. Math. Soc. Transl. (2) **132** (1986), 15–31, Amer. Math. Soc., Providence, RI. MR Zbl

[Kostochka 1984] A. V. Kostochka, "Lower bound of the Hadwiger number of graphs by their average degree", *Combinatorica* **4**:4 (1984), 307–316. MR Zbl

[Kupavskii 2016] A. Kupavskii, "On random subgraphs of Kneser and Schrijver graphs", *J. Combin. Theory Ser. A* **141** (2016), 8–15. MR Zbl

[Lovász 1978] L. Lovász, "Kneser's conjecture, chromatic number, and homotopy", *J. Combin. Theory Ser. A* **25**:3 (1978), 319–324. MR Zbl

[Seymour 2016] P. Seymour, "Hadwiger's conjecture", pp. 417–437 in *Open problems in mathematics*, edited by J. F. Nash, Jr. and M. T. Rassias, Springer, 2016. MR Zbl

[Valencia-Pabon and Vera 2005] M. Valencia-Pabon and J.-C. Vera, "On the diameter of Kneser graphs", *Discrete Math.* **305**:1-3 (2005), 383–385. MR Zbl

Received: 2018-10-31 Revised: 2019-02-19 Accepted: 2019-05-11

hamma@winthrop.edu *Department of Mathematics, Winthrop University, Rock Hill, SC, United States*

meltonkh@miamioh.edu *Department of Mathematics, Miami University, Oxford, OH, United States*

INVOLVE 12:7 (2019)
dx.doi.org/10.2140/involve.2019.12.1163

A binary unrelated-question RRT model accounting for untruthful responding

Amber Young, Sat Gupta and Ryan Parks

(Communicated by Javier Rojo)

Estimating the prevalence of a sensitive trait in a population is not a simple task due to the general tendency among survey respondents to answer sensitive questions in a way that is socially desirable. Use of randomized response techniques (RRT) is one of several approaches for reducing the impact of this tendency. However, despite the additional privacy provided by RRT models, some respondents may still provide an untruthful response. We consider the impact of untruthful responding on binary unrelated-question RRT models and observe that even if only a small number of respondents lie, a significant bias may be introduced to the model. We propose a binary unrelated-question model that accounts for those respondents who may respond untruthfully. This adds an extra layer of precaution to the estimation of the sensitive trait and decreases the importance of presurvey respondent training. Our results are validated using a simulation study.

1. Introduction

Social desirability bias (SDB) refers to the tendency among survey respondents to answer sensitive questions in a way that is viewed positively by others. SDB can interfere with estimation of the prevalence of a sensitive trait in a given population due to potential untruthful responding. There have been many methods proposed to correct for SDB such as the bogus pipeline method introduced by [Jones and Sigall 1971] and the modified Marlowe–Crowne social desirability scale investigated by [Reynolds 1982]. Here, we will focus on another method of reducing the impact of SDB — the randomized response technique (RRT).

RRT was originally introduced by [Warner 1965] and has since been generalized by many researchers including [Greenberg et al. 1969; Gupta et al. 2002; 2013; Christofides 2003; Kim et al. 2006; Nayak and Adeshiyan 2009; Barabesi and

MSC2010: 62D05.
Keywords: model efficiency, optional randomized response models, unrelated-question RRT model, untruthful responding.

Marcheselli 2010; Suarez and Gupta 2018]. This method allows respondents to provide a scrambled response to a sensitive question, where responses cannot be unscrambled at an individual level. This increases respondent privacy and encourages survey participants to respond honestly. Specifically, we will focus on the unrelated-question RRT model introduced in [Greenberg et al. 1969] and the variation of this model developed in [Sihm et al. 2016].

In the [Greenberg et al. 1969] model, a randomization device contains two questions — the sensitive question and an unrelated question. The respondent uses the randomization device and responds to whichever question they receive. The researcher does not know which question each individual respondent answered, but the proportion of respondents with the sensitive trait can still be estimated at an aggregate level. In this model, every respondent uses the randomization device.

However, a topic that is sensitive to one respondent may not be sensitive to another. Optional RRT models, introduced by [Gupta et al. 2002], allow those respondents who do not find the question sensitive to answer the sensitive question directly rather than use the randomization device. The researcher does not know whether the respondent answered directly or used the randomization device.

A variation of the unrelated-question RRT model where the use of the randomization device is optional was introduced in [Gupta et al. 2013]. Respondents are instructed to answer using the randomization device if they find the question sensitive, or respond directly to the sensitive question if they do not find it sensitive. They proposed both quantitative and binary models, and used split sampling in each case to estimate the prevalence of the sensitive trait as well as the sensitivity level of the question.

Binary and quantitative optional unrelated-question models that use a two-question approach to estimating the prevalence of the sensitive trait and the sensitivity level of the question were then developed in [Sihm et al. 2016]. The first question uses the original model of [Greenberg et al. 1969] to estimate the sensitivity level of the question, and the second question uses an optional model similar to that in [Gupta et al. 2013] to estimate the prevalence of the sensitive trait.

However, all of the aforementioned RRT models make the assumption of completely truthful responding. When this assumption is not met, a bias is introduced into these models. This can be a dangerous when asking a very sensitive question, or when the proper presurvey respondent training has not been performed because, in these situations, the proportion of untruthful responding could be relatively large. We assume that the reason for untruthful responding is a respondent's lack of trust in the randomization process — i.e., the belief that the randomization device does not completely protect their privacy.

In this paper, we propose a two-question model that accounts for untruthful responding. The first question uses the [Greenberg et al. 1969] model to estimate

the proportion of respondents who trust the randomization process. The second question asks the respondents to respond using a second randomization device if they trust the randomization process, or simply answer the unrelated question if they do not. This diverts the untruthful responders from introducing bias to the estimation of the sensitive trait. Comparisons of mean squared error are used to demonstrate in which situations this model may be preferred over the models of [Greenberg et al. 1969] and [Sihm et al. 2016].

2. Background binary RRT models

2.1. The original unrelated-question RRT model was introduced in [Greenberg et al. 1969]. In this model, each respondent in a simple random sample with replacement (SRSWR) uses a randomization device and receives the sensitive or unrelated question with probabilities p or $1-p$ respectively. The prevalence of the sensitive characteristic π_x is unknown and the prevalence of the unrelated characteristic π_y is assumed to be known. Recall that in the original Greenberg model, the probability of a "yes" response is given by

$$P_y = p\pi_x + (1-p)\pi_y. \tag{2-1}$$

Rearranging for π_x, we get

$$\pi_x = \frac{P_y - (1-p)\pi_y}{p} := \pi_{\mathrm{GR}}, \tag{2-2}$$

which leads to the estimator

$$\hat{\pi}_{\mathrm{GR}} = \frac{\widehat{P}_y - (1-p)\pi_y}{p}, \tag{2-3}$$

where $\widehat{P}_y$ is the proportion of "yes" responses in the survey. The estimator variance is given by

$$\mathrm{Var}(\hat{\pi}_{\mathrm{GR}}) = \frac{P_y(1-P_y)}{np^2}. \tag{2-4}$$

2.2. A binary optional unrelated-question model using a two-question approach to estimate prevalence of the sensitive trait in a population and the sensitivity level of the trait was proposed in [Sihm et al. 2016].

In this model, respondents are asked two questions. The first question is the research question of interest, and is answered using the model of [Gupta et al. 2013]. The second question uses the [Greenberg et al. 1969] model to ask whether the respondent finds the main research question sensitive. In Question 1, π_x is the known prevalence of the sensitive trait, p is the probability of a respondent receiving the sensitive question, π_y is the prevalence of some unrelated characteristic, and W is the proportion of respondents who find the question sensitive. In Question 2, W

is again the sensitivity level, π_b is the known prevalence of a different unrelated trait, and p_b is the probability of a respondent receiving the question about whether they find the question of interest sensitive.

The probability of a "yes" response for Questions 1 and 2 respectively is represented as

$$P_{y_1} = W[p\pi_x + (1-p)\pi_y] + (1-W)\pi_x, \tag{2-5}$$

$$P_{y_2} = p_b W + (1-p_b)\pi_b. \tag{2-6}$$

Solving (2-6) for W and (2-5) for π_x we have

$$W = \frac{P_{y_2} - (1-p_b)\pi_b}{p_b} \quad \text{and} \quad \pi_x = \frac{P_{y_1} - (1-p)W\pi_y}{1-(1-p)W} := \pi_{\mathrm{SI}}. \tag{2-7}$$

This leads to the estimators

$$\widehat{W} = \frac{\widehat{P}_{y_2} - (1-p_b)\pi_b}{p_b}, \tag{2-8}$$

$$\hat{\pi}_{\mathrm{SI}} = \frac{\widehat{P}_{y_1} - (1-p)\widehat{W}\pi_y}{1-(1-p)\widehat{W}}. \tag{2-9}$$

Using a first-order Taylor approximation for $\hat{\pi}_{\mathrm{SI}}$, the variance of this estimator becomes

$$\mathrm{Var}(\hat{\pi}_{\mathrm{SI}}) \approx \frac{P_{y_1}(1-P_{y_1})}{n(1-(1-p)W)^2} + \frac{(1-p)^2(P_{y_1}-\pi_y)^2 P_{y_2}(1-P_{y_2})}{np_b^2(1-(1-p)W)^4}. \tag{2-10}$$

3. The effect of lying on existing binary unrelated question RRT models

3.1. We first consider [Greenberg et al. 1969]. Let π_a represent the probability that a respondent who has the sensitive trait (belongs to the π_x group) will give a truthful response when confronted with a question about their possession of that trait. It is assumed that those who receive the unrelated question will always provide a truthful response.

Therefore, the probability of a "yes" response in (2-1) becomes

$$P_y^* = p\pi_x\pi_a + (1-p)\pi_y, \tag{3-1}$$

and the estimate in (2-3), mistakenly assuming $\pi_a = 1$, becomes

$$\hat{\pi}_{\mathrm{GR}}^* = \frac{\widehat{P}_y^* - (1-p)\pi_y}{p}, \tag{3-2}$$

with variance

$$\mathrm{Var}(\hat{\pi}_{\mathrm{GR}}^*) = \frac{P_y^*(1-P_y^*)}{np^2}. \tag{3-3}$$

The bias in this estimate is

$$\text{Bias}(\hat{\pi}^*_{\text{GR}}) = \text{E}[\hat{\pi}^*_{\text{GR}} - \pi_x] = \pi_x(\pi_a - 1), \tag{3-4}$$

and therefore the mean squared error of the estimate is

$$\text{MSE}(\hat{\pi}^*_{\text{GR}}) = \text{Var}(\hat{\pi}^*_{\text{GR}}) + \text{Bias}^2(\hat{\pi}^*_{\text{GR}}) = \frac{P^*_y(1-P^*_y)}{np^2} + \pi_x^2(\pi_a - 1)^2. \tag{3-5}$$

3.2. We now consider [Sihm et al. 2016]. Again, let π_a represent the probability of a truthful response as described in Section 3.1. In this model, we assume that respondents who do not find the question sensitive will be honest about their possession of the trait. We also assume that there will be no dishonesty in responses to either question in Question 2, or in response to the unrelated question in Question 1. The probability of a "yes" from (2-5) becomes

$$P^*_{y_1} = W[p\pi_x\pi_a + (1-p)\pi_y] + (1-W)\pi_x, \tag{3-6}$$

and the estimate in (2-9), mistakenly assuming $\pi_a = 1$, becomes

$$\hat{\pi}^*_{\text{SI}} = \frac{\widehat{P}^*_{y_1} - (1-p)\widehat{W}\pi_y}{1-(1-p)\widehat{W}}. \tag{3-7}$$

The first-order Taylor approximation of this estimator is

$$\hat{\pi}^*_{\text{SI}} \approx \frac{P^*_{y_1} - W(1-p)\pi_y}{1-(1-p)W} + \frac{\widehat{P}^*_{y_1} - P^*_{y_1}}{1-(1-p)W} + \frac{(1-p)(P^*_{y_1} - \pi_y)(\widehat{W} - W)}{(1-(1-p)W)^2}, \tag{3-8}$$

which has an approximate variance of

$$\text{Var}(\hat{\pi}^*_{\text{SI}}) \approx \frac{P^*_{y1}(1-P^*_{y_1})}{n(1-(1-p)W)^2} + \frac{(1-p)^2(P^*_{y_1} - \pi_y)^2 P^*_{y_2}(1-P^*_{y_2})}{np_b^2(1-(1-p)W)^4}. \tag{3-9}$$

The bias for the estimate in (3-8) is

$$\text{Bias}(\hat{\pi}^*_{\text{SI}}) \approx \text{E}[\hat{\pi}^*_{\text{SI}} - \pi_x] = \frac{W\pi_x p(\pi_a - 1)}{1-(1-p)W}, \tag{3-10}$$

and therefore the mean squared error of the estimate is

$$\begin{aligned}
&\text{MSE}(\hat{\pi}^*_{\text{SI}}) \\
&= \text{Var}(\hat{\pi}^*_{\text{SI}}) + \text{Bias}^2(\hat{\pi}^*_{\text{SI}}) \\
&= \frac{P^*_{y_1}(1-P^*_{y_1})}{n(1-(1-p)W)^2} + \frac{(1-p)^2(P^*_{y_1} - \pi_y)^2 P^*_{y_2}(1-P^*_{y_2})}{np_b^2(1-(1-p)W)^4} + \left(\frac{W\pi_x p(\pi_a - 1)}{1-(1-p)W}\right)^2.
\end{aligned} \tag{3-11}$$

4. The proposed model

The goal of this model is to avoid any bias introduced to the model by untruthful responding. To do this, we propose a two-question model where the first question uses the [Greenberg et al. 1969] model to ask whether respondents trust the randomization process. For the second question, respondents are asked to respond using the [Greenberg et al. 1969] model if they trust the randomization process, or simply respond to the unrelated question if they do not. This way, anyone who may be tempted to provide an untruthful answer about their involvement in the sensitive question of interest is redirected to the unrelated question.

Let π_x be the prevalence of the sensitive trait of interest, π_y be the known prevalence of some unrelated trait, π_b be the known prevalence of some other unrelated trait, p_b be the probability of receiving the question about trust in Question 1, and p be the probability of receiving the sensitive question in Question 2. Also, let π_a be the probability that a respondent will trust the randomization process (the probability someone would not give an untruthful response when faced with the sensitive question).

The probability of a "yes" response to Question i ($i = 1, 2$) is represented as

$$P_{y_1} = p_b\pi_a + (1 - p_b)\pi_b, \tag{4-1}$$

$$P_{y_2} = \pi_a[p\pi_x + (1 - p)\pi_y] + (1 - \pi_a)\pi_y. \tag{4-2}$$

Solving (4-1) and (4-2) for π_a and π_x gives us

$$\pi_a = \frac{P_{y_1} - (1 - p_b)\pi_b}{p_b} \quad \text{and} \quad \pi_x = \frac{P_{y_2} - \pi_y(1 - \pi_a p)}{\pi_a p}, \tag{4-3}$$

which leads to the estimates

$$\hat{\pi}_a = \frac{\widehat{P}_{y_1} - (1 - p_b)\pi_b}{p_b} \quad \text{and} \quad \hat{\pi}_x = \frac{\widehat{P}_{y_2} - \pi_y(1 - \hat{\pi}_a p)}{\hat{\pi}_a p}, \tag{4-4}$$

where $\widehat{P}_{y_i}$ is the proportion of respondents who respond "yes" to Question i ($i = 1, 2$).

Observe that $\hat{\pi}_a$ is an unbiased estimator of π_a and its variance is

$$\mathrm{Var}(\hat{\pi}_a) = \frac{P_{y_1}(1 - P_{y_1})}{np_b}. \tag{4-5}$$

Using a first-order Taylor approximation for $\hat{\pi}_x$ gives us

$$\hat{\pi}_x \approx \frac{P_{y_2} - \pi_y(1 - \pi_y p)}{\pi_a p} + \frac{\widehat{P}_{y_2} - P_{y_2}}{\pi_a p} + \frac{p(\pi_y - P_{y_2})(\hat{\pi}_a - \pi_a)}{(\pi_a p)^2} := \hat{\pi}_{\mathrm{YO}}. \tag{4-6}$$

The estimate $\hat{\pi}_{\mathrm{YO}}$ is an unbiased estimator of π_x up to a first-order Taylor approximation, and its variance is given by

$$\mathrm{Var}(\hat{\pi}_{\mathrm{YO}}) = \frac{P_{y_2}(1 - P_{y_2})}{n(\pi_a p)^2} + \frac{P_{y_1}(1 - P_{y_1})p^2(\pi_y - P_{y_2})^2}{np_b^2(\pi_a p)^4}. \tag{4-7}$$

5. Simulation results

We now present simulation results for our estimator $\hat{\pi}_{\mathrm{YO}}$ and compare it to the estimators $\hat{\pi}^*_{\mathrm{GR}}$ and $\hat{\pi}^*_{\mathrm{SI}}$ as detailed in Sections 3.1 and 3.2, respectively. Table 1 details the simulation results using 10,000 iterations at $n = 500$. We allow π_a (the

			π_a					
			1.00	0.99	0.95	0.90	0.85	0.80
proposed model		$\hat{\pi}_{\mathrm{YO}}$	0.199906	0.199673	0.199971	0.200081	0.199688	0.200349
		$\widehat{\mathrm{Var}}(\hat{\pi}_{\mathrm{YO}})$	0.000540	0.000548	0.000602	0.000684	0.000778	0.000879
		$\mathrm{Var}(\hat{\pi}_{\mathrm{YO}})$	0.000541	0.000553	0.000608	0.000686	0.000780	0.000891
		$\hat{\pi}_a$	0.999841	0.989917	0.949589	0.900047	0.850556	0.799957
		$\widehat{\mathrm{Var}}(\hat{\pi}_a)$	0.000450	0.000479	0.000536	0.000599	0.000666	0.000690
		$\mathrm{Var}(\hat{\pi}_a)$	0.000461	0.000477	0.000536	0.000601	0.000656	0.000701
simple unrelated		$\hat{\pi}^*_{\mathrm{GR}}$	0.199769	0.198214	0.189576	0.180255	0.170079	0.160274
		$\widehat{\mathrm{Var}}(\hat{\pi}^*_{\mathrm{GR}})$	0.000540	0.000517	0.000517	0.000507	0.000489	0.000473
		$\mathrm{Var}(\hat{\pi}^*_{\mathrm{GR}})$	0.000536	0.000533	0.000522	0.000507	0.000492	0.000477
		$\mathrm{Bias}(\hat{\pi}^*_{\mathrm{GR}})$	0.000000	0.002000	0.010000	0.020000	0.030000	0.040000
		$\mathrm{MSE}(\hat{\pi}^*_{\mathrm{GR}})$	0.000536	0.000537	0.000622	0.000907	0.001392	0.002077
optional two-question	$W = 0.70$	$\hat{\pi}^*_{\mathrm{SI}}$	0.200002	0.198722	0.193519	0.187040	0.180684	0.174260
		$\widehat{\mathrm{Var}}(\hat{\pi}^*_{\mathrm{SI}})$	0.000457	0.000440	0.000444	0.000431	0.000427	0.000414
		$\mathrm{Var}(\hat{\pi}^*_{\mathrm{SI}})$	0.000455	0.000454	0.000447	0.000438	0.000429	0.000420
		$\mathrm{Bias}(\hat{\pi}^*_{\mathrm{SI}})$	0.000000	0.001302	0.006512	0.013023	0.019535	0.026047
		$\mathrm{MSE}(\hat{\pi}^*_{\mathrm{SI}})$	0.000455	0.000455	0.000489	0.000607	0.000810	0.001098
	$W = 0.80$	$\hat{\pi}^*_{\mathrm{SI}}$	0.199866	0.198278	0.191730	0.184722	0.176871	0.169212
		$\widehat{\mathrm{Var}}(\hat{\pi}^*_{\mathrm{SI}})$	0.000487	0.000487	0.000456	0.000461	0.000448	0.000435
		$\mathrm{Var}(\hat{\pi}^*_{\mathrm{SI}})$	0.000480	0.000478	0.000470	0.000459	0.000449	0.000438
		$\mathrm{Bias}(\hat{\pi}^*_{\mathrm{SI}})$	0.000000	0.001524	0.007619	0.015238	0.022857	0.030476
		$\mathrm{MSE}(\hat{\pi}^*_{\mathrm{SI}})$	0.000480	0.000481	0.000528	0.000692	0.000971	0.001366
	$W = 0.90$	$\hat{\pi}^*_{\mathrm{SI}}$	0.200340	0.198214	0.191183	0.182637	0.173468	0.165007
		$\widehat{\mathrm{Var}}(\hat{\pi}^*_{\mathrm{SI}})$	0.000501	0.000510	0.000505	0.000485	0.000482	0.000452
		$\mathrm{Var}(\hat{\pi}^*_{\mathrm{SI}})$	0.000507	0.000505	0.000495	0.000483	0.000470	0.000457
		$\mathrm{Bias}(\hat{\pi}^*_{\mathrm{SI}})$	0.000000	0.001756	0.008780	0.017561	0.026341	0.035122
		$\mathrm{MSE}(\hat{\pi}^*_{\mathrm{SI}})$	0.000507	0.000508	0.000572	0.000791	0.001164	0.001690

Table 1. Simulation results for all models under untruthful responding: iterations $= 10,000$, $n = 500$, $p = 0.8$, $p_b = 0.8$, $\pi_y = 0.3$, $\pi_b = 0.1$, $\pi_x = 0.2$.

π_a	π_x		
	0.10	0.20	0.30
1.00	0.9533	0.9915	**1.0000**
0.99	0.9224	0.9711	0.9890
0.95	0.8519	**1.0234**	**1.1907**
0.90	0.8519	**1.3219**	**1.8816**
0.85	0.9117	**1.7858**	**2.8979**
0.80	**1.0003**	**2.3302**	**4.0853**

Table 2. Percent relative efficiency $\text{PRE}(\hat{\pi}_{\text{YO}}, \hat{\pi}^*_{\text{GR}})$ under untruthful responding: $n = 500,\ p = 0.8,\ p_b = 0.8,\ \pi_y = 0.3,\ \pi_b = 0.1$.

proportion of truthful responding) to vary and fix other parameters at $\pi_x = 0.2$, $p = 0.8$, $p_b = 0.8$, $\pi_y = 0.3$, and $\pi_b = 0.1$. Note that the proposed model's estimate for the proportion of truthful responding ($\hat{\pi}_a$) is also included in Table 1.

Notice that $\hat{\pi}^*_{\text{GR}}$ and $\hat{\pi}^*_{\text{SI}}$ underestimate the prevalence of the sensitive trait when the proportion of truthful responding is less than 1. This is due to the bias introduced to these models under untruthful responding. To compare the efficiency of the proposed model to those of existing binary RRT models when some untruthful responding is suspected, we use the *percent relative efficiency* (PRE), where

$$\text{PRE}(\hat{\pi}_{\text{YO}}, \hat{\pi}^*_{\text{GR}}) = \frac{\text{MSE}(\hat{\pi}^*_{\text{GR}})}{\text{MSE}(\hat{\pi}_{\text{YO}})}, \tag{5-1}$$

$$\text{PRE}(\hat{\pi}_{\text{YO}}, \hat{\pi}^*_{\text{SI}}) = \frac{\text{MSE}(\hat{\pi}^*_{\text{SI}})}{\text{MSE}(\hat{\pi}_{\text{YO}})}. \tag{5-2}$$

A PRE value of 1 or greater favors the proposed model over the existing model. The proposed model is unbiased under untruthful responding, therefore

$$\text{MSE}(\hat{\pi}_{\text{YO}}) = \text{Var}(\hat{\pi}_{\text{YO}}). \tag{5-3}$$

The comparison of the proposed model with the [Greenberg et al. 1969] model with untruthful responding can be seen in Table 2. We can see that when the prevalence of the sensitive trait is at least 20% ($\pi = 0.20$), and as low as only 5% of respondents give untruthful responses ($\pi_a = 0.95$), the proposed model is generally preferred over the original Greenberg model.

The comparison of the proposed model to that of [Sihm et al. 2016] can be found in Table 3. We can see that when the sensitivity level of the question, the prevalence of the sensitive trait, or the proportion of respondents who give an untruthful response increases, the proposed model tends to be more efficient than that of [Sihm et al. 2016].

W	π_a	π_x		
		0.10	0.20	0.30
0.70	1.00	0.7689	0.8417	0.8653
	0.99	0.7439	0.8226	0.8511
	0.95	0.6700	0.8045	0.8994
	0.90	0.6224	0.8847	**1.1486**
	0.85	0.6069	**1.0392**	**1.5386**
	0.80	0.6093	**1.2318**	**2.0041**
0.80	1.00	0.8268	0.8882	0.9070
	0.99	0.8000	0.8685	0.8935
	0.95	0.7260	0.8688	0.9829
	0.90	0.6901	**1.0075**	**1.3514**
	0.85	0.6942	**1.2454**	**1.9110**
	0.80	0.7192	**1.5330**	**2.5721**
0.90	1.00	0.8891	0.9382	0.9518
	0.99	0.8603	0.9181	0.9394
	0.95	0.7873	0.9416	**1.0794**
	0.90	0.7671	**1.1525**	**1.5940**
	0.85	0.7959	**1.4927**	**2.3607**
	0.80	0.8493	**1.8966**	**3.2605**

Table 3. Percent relative efficiency $\mathrm{PRE}(\hat{\pi}_{\mathrm{YO}}, \hat{\pi}^*_{\mathrm{SI}})$ under untruthful responding: $n = 500,\ p = 0.8,\ p_b = 0.8,\ \pi_y = 0.3,\ \pi_b = 0.1$.

6. Conclusion

We propose a binary unrelated-question RRT model that accounts for untruthful responding. This model provides an unbiased estimator, whereas existing models are biased under untruthful responding. This provides an additional layer of precaution to the estimation of a sensitive trait. We found that there are many scenarios in which this model would be preferred over the model of [Greenberg et al. 1969] and even when it would be preferred over an optional binary RRT model as in [Sihm et al. 2016].

For instance, when the prevalence of the sensitive trait is high or the proportion of untruthful responding is high, we found that the proposed model has a higher efficiency than that of [Greenberg et al. 1969] and [Sihm et al. 2016]. However, when the proportion of untruthful responding is low and the prevalence of the sensitive trait is also low, it may not be worth expending the extra energy in estimating π_a. It also may not be worth expending the energy when comparing the proposed model to [Sihm et al. 2016] when the sensitivity level of the question is low.

In examining the advantage of the proposed method over the existing methods, we have relied on the commonly used approach of looking at the percent relative

π_a	π_x	
	0.20	0.30
1.00	587	546
0.99	599	552
0.95	558	451
0.90	415	277
0.85	298	177
0.80	224	125

Table 4. Sample size needed for proposed model to achieve the same efficiency as Greenberg model under untruthful responding ($\text{MSE}(\hat{\pi}^*_{\text{GR}})$) with a fixed sample size of $n=500,\ p=0.8,\ p_b=0.8,$ $\pi_y=0.3,\ \pi_b=0.1$.

efficiency, as seen in Tables 2 and 3. In this approach, we keep the sample size fixed and look at the mean squared error (MSE) of one model as compared to the other. An alternative approach could be to look at the MSE of one model with a fixed sample size, and then see what sample size would be necessary for the proposed model to achieve the same efficiency. Limited results are presented in Table 4 to give the reader some idea as to how much reduction in sample size can be achieved by the proposed method. It is clear by these results that when the proportion of untruthful responding is high or the prevalence of the sensitive trait is high, the proposed model can offer a large reduction in sample size while achieving the same efficiency of other models. However, when either of these values is very low, it again may not be worth the energy to estimate π_a.

It is also important to note that, because the proposed model is unbiased under untruthful responding, it eliminates the need for extensive presurvey training of respondents, as the purpose of presurvey training is to minimize untruthful responding.

Acknowledgements

The authors would like to express their sincere appreciation for the careful reading of the original version of this paper by the reviewer and the editor. Their comments have helped improve the presentation of our work considerably.

This material is based upon work supported by the National Science Foundation, grant no. DMS-1560332.

References

[Barabesi and Marcheselli 2010] L. Barabesi and M. Marcheselli, "Bayesian estimation of proportion and sensitivity level in randomized response procedures", *Metrika* **72**:1 (2010), 75–88. MR Zbl

[Christofides 2003] T. C. Christofides, "A generalized randomized response technique", *Metrika* **57**:2 (2003), 195–200. MR Zbl

[Greenberg et al. 1969] B. G. Greenberg, A.-L. A. Abul-Ela, W. R. Simmons, and D. G. Horvitz, "The unrelated question randomized response model: theoretical framework", *J. Amer. Statist. Assoc.* **64**:326 (1969), 520–539. MR

[Gupta et al. 2002] S. Gupta, B. Gupta, and S. Singh, "Estimation of sensitivity level of personal interview survey questions", *J. Statist. Plann. Inference* **100**:2 (2002), 239–247. MR Zbl

[Gupta et al. 2013] S. Gupta, A. Tuck, T. Spears Gill, and M. Crowe, "Optional unrelated-question randomized response models", *Involve* **6**:4 (2013), 483–492. MR Zbl

[Jones and Sigall 1971] E. E. Jones and H. Sigall, "The bogus pipeline: a new paradigm for measuring affect and attitude", *Psych. Bull.* **76**:5 (1971), 349–364.

[Kim et al. 2006] J.-M. Kim, J. M. Tebbs, and S.-W. An, "Extensions of Mangat's randomized-response model", *J. Statist. Plann. Inference* **136**:4 (2006), 1554–1567. MR Zbl

[Nayak and Adeshiyan 2009] T. K. Nayak and S. A. Adeshiyan, "A unified framework for analysis and comparison of randomized response surveys of binary characteristics", *J. Statist. Plann. Inference* **139**:8 (2009), 2757–2766. MR Zbl

[Reynolds 1982] W. M. Reynolds, "Development of reliable and valid short forms of the Marlowe–Crowne social desirability scale", *J. Clinic. Psych.* **38**:1 (1982), 119–125.

[Sihm et al. 2016] J. S. Sihm, A. Chhabra, and S. N. Gupta, "An optional unrelated question RRT model", *Involve* **9**:2 (2016), 195–209. MR Zbl

[Suarez and Gupta 2018] D. P. Suarez and S. Gupta, "Variations of the Greenberg unrelated question binary model", *Involve* **11**:1 (2018), 119–126. MR Zbl

[Warner 1965] S. L. Warner, "Randomized response: a survey technique for eliminating evasive answer bias", *J. Amer. Statist. Assoc.* **60**:309 (1965), 63–69. Zbl

Received: 2018-11-02 Revised: 2019-03-23 Accepted: 2019-05-01

young268@purdue.edu *Department of Statistics, Purdue University, West Lafayette, IN, United States*

sngupta@uncg.edu *Department of Mathematics and Statistics, University of North Carolina, Greensboro, NC, United States*

rjparks@uncg.edu *Department of Mathematics and Statistics, University of North Carolina, Greensboro, NC, United States*

INVOLVE 12:7 (2019)
dx.doi.org/10.2140/involve.2019.12.1175

Toward a Nordhaus–Gaddum inequality for the number of dominating sets

Lauren Keough and David Shane

(Communicated by Kenneth S. Berenhaut)

A dominating set in a graph G is a set S of vertices such that every vertex of G is either in S or is adjacent to a vertex in S. Nordhaus–Gaddum inequalities relate a graph G to its complement $\overline{G}$. In this spirit Wagner proved that any graph G on n vertices satisfies $\partial(G) + \partial(\overline{G}) \geq 2^n$, where $\partial(G)$ is the number of dominating sets in a graph G. In the same paper he commented that proving an upper bound for $\partial(G) + \partial(\overline{G})$ among all graphs on n vertices seems to be much more difficult. Here we prove an upper bound on $\partial(G) + \partial(\overline{G})$ and prove that any graph maximizing this sum has minimum degree at least $\lfloor n/2 \rfloor - 2$ and maximum degree at most $\lceil n/2 \rceil + 1$. We conjecture that the complete balanced bipartite graph maximizes $\partial(G) + \partial(\overline{G})$ and have verified this computationally for all graphs on at most 10 vertices.

1. Introduction

A *dominating set* in a graph G is a set of vertices S such that every vertex of G is either in S or adjacent to a vertex in S. Dominating sets, and their many variations, have long been studied [Haynes et al. 1998]. Also long-studied are Nordhaus–Gaddum inequalities, which describe the relationship between a graph parameter on G and the same graph parameter on $\overline{G}$, the complement of G, in terms of the order of the graph. The original Nordhaus–Gaddum inequalities concern the chromatic number of a graph G, denoted by $\chi(G)$. Nordhaus and Gaddum [1956] proved that if G has n vertices then

$$2\sqrt{n} \leq \chi(G) + \chi(\overline{G}) \leq n + 1$$

and

$$n \leq \chi(G) \cdot \chi(\overline{G}) \leq \left(\frac{n+1}{2}\right)^2.$$

Since then there have been several hundred papers proving similar relations for many different graph parameters [Aouchiche and Hansen 2013]. In particular, there

MSC2010: 05C35, 05C69.
Keywords: Nordhaus–Gaddum inequalities, dominating sets.

are such inequalities for the domination number (the size of a smallest dominating set) [Jaeger and Payan 1972; Borowiecki 1976]. See [Aouchiche and Hansen 2013] and [Harary and Haynes 1996] for surveys of results concerning Nordhaus–Gaddum inequalities for at least 30 types of domination numbers.

Separately, there has been interest in results concerning maximizing or minimizing the *number* of a given graph substructure, rather than its size, subject to certain conditions. For a survey on these types of problems for regular graphs see [Zhao 2017]. Recently, there have been several papers that maximize or minimize the total number of dominating sets or total dominating sets for connected graphs of a given order [Bród and Skupień 2006; Wagner 2013; Skupień 2014; Krzywkowski and Wagner 2018].

Let $\partial(G)$ be the number of dominating sets in a graph G. Uniting the ideas of Nordhaus–Gaddum inequalities and counting the number of graph substructures, Wagner [2013] proved that

$$\partial(G) + \partial(\overline{G}) \geq 2^n.$$

In the same paper, he proposed that determining the maximum of $\partial(G) + \partial(\overline{G})$ as G ranges over all possible graphs on n vertices seems to be much more difficult. We are able to prove the following theorem.

Theorem 1.1. *If G is a graph on n vertices, then*

$$\partial(G) + \partial(\overline{G}) \leq 2^{n+1} - 2^{\lfloor n/2 \rfloor} - 2^{\lceil n/2 \rceil - 1}.$$

However, this is not the least upper bound. The authors and Wagner conjecture that the extremal graph is the complete balanced bipartite graph, leading to the following conjecture.

Conjecture 1.2. *For a graph G on n vertices,*

$$\begin{aligned}\partial(G) + \partial(\overline{G}) &\leq 2(2^{\lfloor n/2 \rfloor} - 1)(2^{\lceil n/2 \rceil} - 1) + 2 \\ &= \partial(K_{\lfloor n/2 \rfloor, \lceil n/2 \rceil}) + \partial(\overline{K}_{\lfloor n/2 \rfloor, \lceil n/2 \rceil}).\end{aligned}$$

This conjecture has been verified up to $n = 10$ vertices. Wagner pointed out that this conjecture makes heuristic sense as both the complete balanced bipartite graph and its complement can be dominated by only two vertices (personal communication, October 3, 2017).

Throughout the paper we use $N_G(v)$ to mean the open neighborhood of the vertex v in the graph G and $N_G[v]$ for the closed neighborhood of v in G. If S is a set of vertices we define $N_G(S)$ and $N_G[S]$ similarly. In Section 2 we prove Theorem 1.1. In Section 3 we provide a maximum and minimum degree condition for the extremal graph. Finally, in Section 4 we provide some asymptotics and describe some of the difficulties in finding the least upper bound for $\partial(G) + \partial(\overline{G})$.

2. An upper bound for $\partial(G) + \partial(\overline{G})$

To prove that $\partial(G) + \partial(\overline{G}) \geq 2^n$, Wagner [2013] used the fact that if a set S does not dominate G, then $\overline{S}$ dominates $\overline{G}$. We use this same fact to express the sum of the number of dominating sets in G and $\overline{G}$ as

$$\partial(G) + \partial(\overline{G}) = 2^n + \Upsilon(G, \overline{G}),$$

where

$$\Upsilon(G, \overline{G}) = |\{A \subseteq V(G) : A \text{ dominates } G \text{ and } \overline{A} \text{ dominates } \overline{G}\}|.$$

We make use of $\Upsilon(G, \overline{G})$ to establish the following upper bound.

Lemma 2.1. *If G is a graph on n vertices and a vertex $v \in V(G)$ has $\deg_G(v) = k$, then*

$$\partial(G) + \partial(\overline{G}) \leq 2^{n+1} - 2^k - 2^{n-k-1}.$$

Proof. We bound $\Upsilon(G, \overline{G})$ in terms of n and k and thus bound $\partial(G) + \partial(\overline{G})$ in terms of n and k. It will be helpful to visualize G and $\overline{G}$ as shown in Figure 1. Note that the graphs in Figure 1 do not include any edges that are not incident with v, but every edge is in either G or $\overline{G}$.

Let's consider a set $S \subseteq V(G)$ with the following properties:

- $v \in S$.
- $N_{\overline{G}}(v) = \overline{N_G[v]} \subseteq S$.

We claim that $\overline{S}$ is not a dominating set of $\overline{G}$. Since $\overline{S} \cap N_{\overline{G}}(v) = \varnothing$ and $v \notin \overline{S}$, we have that $v \notin N_{\overline{G}}[S]$. Thus, $\overline{S}$ is not a dominating set of $\overline{G}$. Therefore all sets satisfying the construction of S are not counted in $\Upsilon(G, \overline{G})$. Since each element of $N_G(v)$ may or may not be included in S and $|N_G(v)| = \deg_G(v) = k$, we have identified 2^k sets that are not in $\Upsilon(G, \overline{G})$.

Let's now consider a set $T \subseteq V(G)$ with the following properties:

- $v \notin T$.
- $T \cap N_G(v) = \varnothing$.

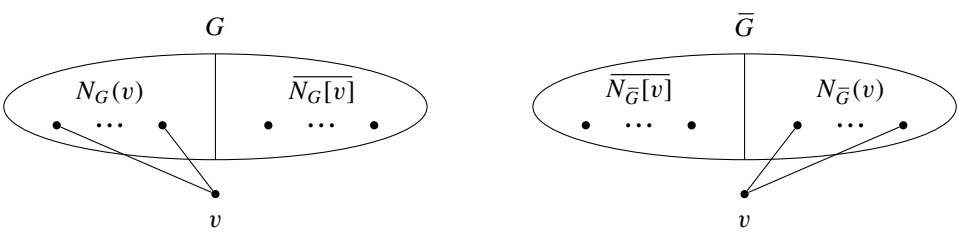

Figure 1. A drawing of G and $\overline{G}$ to aid in the proof of Lemma 2.1.

Since $v \notin N_G[T]$, we know T is not a dominating set of G and all sets satisfying the construction of T are not counted in $\Upsilon(G, \overline{G})$. Since each element of $N_{\overline{G}}(v)$ may or may not be included in T and $|N_{\overline{G}}(v)| = n-k-1$, we have identified 2^{n-k-1} sets that are not in $\Upsilon(G, \overline{G})$.

No sets satisfy the construction of both S and T since $v \in S$ and $v \notin T$ and so we have $2^k + 2^{n-k-1}$ sets that are not counted in $\Upsilon(G, \overline{G})$. We conclude $\Upsilon(G, \overline{G}) \leq 2^n - (2^k + 2^{n-k-1})$ and thus

$$\partial(G) + \partial(\overline{G}) = 2^n + \Upsilon(G, \overline{G}) \leq 2^{n+1} - 2^k - 2^{n-k-1}. \qquad \square$$

To prove Theorem 1.1 we apply Lemma 2.1 for a vertex of degree at least $\lfloor n/2 \rfloor$, which must exist in either G or $\overline{G}$. This eliminates the need for the knowledge of the degree of a specific vertex in G.

Proof of Theorem 1.1. Let G be a graph on n vertices. Since $\max\{\Delta(G), \Delta(\overline{G})\} \geq \lfloor n/2 \rfloor$, there exists some vertex $v \in V(G)$ such that $\deg_G(v) = \lfloor n/2 \rfloor + d$ or $\deg_{\overline{G}}(v) = \lfloor n/2 \rfloor + d$, where $d \geq 0$. Without loss of generality suppose $\deg_G(v) = \lfloor n/2 \rfloor + d$, where $d \geq 0$. From Lemma 2.1 we have

$$\partial(G) + \partial(\overline{G}) \leq 2^{n+1} - 2^{\lfloor n/2 \rfloor + d} - 2^{n-(\lfloor n/2 \rfloor + d)-1} = 2^{n+1} - 2^d \cdot 2^{\lfloor n/2 \rfloor} - \frac{2^{\lceil n/2 \rceil - 1}}{2^d}.$$

Considering the cases $d = 0$ and $d > 0$ separately we have

$$\partial(G) + \partial(\overline{G}) \leq 2^{n+1} - 2^d \cdot 2^{\lfloor n/2 \rfloor} - \frac{2^{\lceil n/2 \rceil - 1}}{2^d} \leq 2^{n+1} - 2^{\lfloor n/2 \rfloor} - 2^{\lceil n/2 \rceil - 1}. \qquad \square$$

3. Degree condition

We now use Lemma 2.1 and our conjectured extremal graph to get a degree condition on all possible extremal graphs.

Theorem 3.1. *If G is a graph on n vertices that maximizes $\partial(G) + \partial(\overline{G})$, then $\min\{\delta(G), \delta(\overline{G})\} \geq \lfloor n/2 \rfloor - 2$ and $\max\{\Delta(G), \Delta(\overline{G})\} \leq \lceil n/2 \rceil + 1$.*

Proof. Let G be a graph on n vertices such that G maximizes $\partial(G) + \partial(\overline{G})$. First suppose n is even. Suppose that for some $v \in V(G)$, we have $\deg_G(v) \geq n/2 + d$ for some integer $d \geq 2$. By Lemma 2.1,

$$\begin{aligned}
\partial(G) + \partial(\overline{G}) &\leq 2^{n+1} - 2^{n/2+d} - 2^{n-(n/2+d)-1} \\
&= 2^{n+1} - 2^{d-1} \cdot 2^{n/2+1} - \frac{2^{n/2+1}}{2^{d+2}} \\
&< 2^{n+1} - 2 \cdot 2^{n/2+1} \\
&< 2^{n+1} - 2^{n/2+2} + 4 \\
&= \partial(K_{n/2,n/2}) + \partial(\overline{K}_{n/2,n/2}).
\end{aligned}$$

This contradicts that G is extremal. Therefore, $\deg_G(v) \le n/2+1$. The same argument applies for $\overline{G}$, so $\deg_{\overline{G}}(v) \le n/2+1$. For any vertex v, we have $\deg_G(v)+\deg_{\overline{G}}(v) = n-1$ so these upper bounds imply

$$\deg_G(v) \ge n - \left(\frac{n}{2}+1\right) - 1 = \frac{n}{2} - 2,$$
$$\deg_{\overline{G}}(v) \ge n - \left(\frac{n}{2}+1\right) - 1 = \frac{n}{2} - 2.$$

These four inequalities imply the result when n is even.

Now suppose n is odd and that for some $v \in V(G)$, $\deg_G(v) \ge (n+1)/2+d$, where $d \ge 2$. By Lemma 2.1,

$$\begin{aligned}\partial(G)+\partial(\overline{G}) &\le 2^{n+1} - 2^{(n+1)/2+d} - 2^{n-((n+1)/2+d)-1} \\ &= 2^{n+1} - 2^{d-1}\cdot 2^{(n+3)/2} - \frac{2^{(n+1)/2}}{2^{d+2}} \\ &< 2^{n+1} - 2\cdot 2^{(n+3)/2} \\ &< 2^{n+1} - 2^{(n+3)/2} - 2^{(n+1)/2} + 4 \\ &= \partial(K_{(n+1)/2,(n-1)/2}) + \partial(\overline{K}_{(n+1)/2,(n-1)/2}).\end{aligned}$$

Again, this contradicts that G is extremal. Therefore, $\deg_G(v) \le (n+1)/2+1$. As before this implies

$$\deg_{\overline{G}}(v) \le \frac{n+1}{2}+1,$$
$$\deg_G(v) \ge n - \left(\frac{n+1}{2}+1\right) - 1 = \frac{n-1}{2} - 2,$$
$$\deg_{\overline{G}}(v) \ge n - \left(\frac{n+1}{2}+1\right) - 1 = \frac{n-1}{2} - 2,$$

which imply the result when n is odd. □

This theorem could be used in a future proof of Conjecture 1.2, as it eliminates numerous graphs from consideration for each n.

4. Conclusion

There are several obstacles to proving Conjecture 1.2 using some traditional techniques. One strategy would be to start with a graph and move edges between the graph and the complement in a way that increases $\partial(G)+\partial(\overline{G})$ at each edge move. However there are several examples that show this isn't possible. For example, $\partial(C_5)+\partial(\overline{C}_5) = 42$, but moving any edge results in only 40 dominating sets. Using a counting argument one can prove the following.

Proposition 4.1. *For any complete multipartite graph G on n vertices that is not the complete balanced bipartite graph or its complement*

$$\partial(G)+\partial(\overline{G}) < \partial(K_{\lfloor n/2\rfloor,\lceil n/2\rceil}) + \partial(\overline{K}_{\lfloor n/2\rfloor,\lceil n/2\rceil}).$$

If one could show that any extremal graph should be a complete multipartite graph then Proposition 4.1 would complete a proof of Conjecture 1.2.

A proof of Conjecture 1.2 also doesn't work out nicely by induction on the number of vertices. Let H_n be the complete balanced bipartite graph on n vertices, G denote any graph on n vertices and $G+v$ mean the addition of one vertex, v, and any edges we want. We might try to prove that

$$(\partial(H_{n+1})+\partial(H_{n+1}))-(\partial(H_n)+\partial(\overline{H}_n)) > (\partial(G+v)+\partial(\overline{G+v}))-(\partial(G)+\partial(G)).$$

That is, the step from a maximal graph to the maximal graph on one more vertex increases the Nordhaus–Gaddum sum by more than adding a vertex to any other graph would. However, as one example, $G = K_{1,3}$ does not have this property.

Theorem 1.1 does give us a good result asymptotically. To see this, consider how close $\partial(G)+\partial(\overline{G})$ can be to 2^{n+1} (a trivial upper bound). The complete balanced bipartite graph shows that

$$\max\{\partial(G)+\partial(\overline{G})\} \geq 2^{n+1} - 2^{\lfloor n/2\rfloor+1} - 2^{\lceil n/2\rceil+1} + 4,$$

where the maximum is taken over all graphs G on n vertices. This shows that the gap between $\max\{\partial(G)+\partial(\overline{G})\}$ and 2^{n+1} is at most

$$\begin{cases} (4-o(1))\,2^{n/2} & \text{if } n \text{ is even,} \\ (3\sqrt{2}-o(1))\,2^{n/2} & \text{if } n \text{ is odd,} \end{cases}$$

and we conjecture this gap is the smallest possible. From Theorem 1.1 we know

$$\max(\partial(G)+\partial(\overline{G})) \leq 2^{n+1} - 2^{\lfloor n/2\rfloor} - 2^{\lceil n/2\rceil-1},$$

which means that the gap is always at least

$$\begin{cases} \left(\frac{3}{2}\right)2^{n/2} & \text{if } n \text{ is even,} \\ (\sqrt{2})\,2^{n/2} & \text{if } n \text{ is odd.} \end{cases}$$

Therefore, $2^{n/2}$ is the right order of magnitude for the gap between 2^{n+1} and $\max\{\partial(G)+\partial(\overline{G})\}$.

Acknowledgements

Shane was supported by the Alayont Undergraduate Research Fellowship in Mathematics at Grand Valley State University. We would like to thank David Galvin for his contributions to the analysis in the Conclusion and Stefan Wagner for his helpful comments on a draft of this paper.

References

[Aouchiche and Hansen 2013] M. Aouchiche and P. Hansen, "A survey of Nordhaus–Gaddum type relations", *Discrete Appl. Math.* **161**:4-5 (2013), 466–546. MR Zbl

[Borowiecki 1976] M. Borowiecki, "On the external stability number of a graph and its complement", *Prace Nauk. Inst. Mat. Politech. Wrocław.* **12** (1976), 39–43. Zbl

[Bród and Skupień 2006] D. Bród and Z. Skupień, "Trees with extremal numbers of dominating sets", *Australas. J. Combin.* **35** (2006), 273–290. MR Zbl

[Harary and Haynes 1996] F. Harary and T. W. Haynes, "Nordhaus–Gaddum inequalities for domination in graphs", *Discrete Math.* **155**:1-3 (1996), 99–105. MR Zbl

[Haynes et al. 1998] T. W. Haynes, S. T. Hedetniemi, and P. J. Slater, *Fundamentals of domination in graphs*, Monographs and Textbooks in Pure and Appl. Math. **208**, Dekker, New York, 1998. MR Zbl

[Jaeger and Payan 1972] F. Jaeger and C. Payan, "Relations du type Nordhaus–Gaddum pour le nombre d'absorption d'un graphe simple", *C. R. Acad. Sci. Paris Sér. A-B* **274** (1972), 728–730. MR Zbl

[Krzywkowski and Wagner 2018] M. Krzywkowski and S. Wagner, "Graphs with few total dominating sets", *Discrete Math.* **341**:4 (2018), 997–1009. MR Zbl

[Nordhaus and Gaddum 1956] E. A. Nordhaus and J. W. Gaddum, "On complementary graphs", *Amer. Math. Monthly* **63** (1956), 175–177. MR Zbl

[Skupień 2014] Z. Skupień, "Majorization and the minimum number of dominating sets", *Discrete Appl. Math.* **165** (2014), 295–302. MR Zbl

[Wagner 2013] S. Wagner, "A note on the number of dominating sets of a graph", *Util. Math.* **92** (2013), 25–31. MR Zbl

[Zhao 2017] Y. Zhao, "Extremal regular graphs: independent sets and graph homomorphisms", *Amer. Math. Monthly* **124**:9 (2017), 827–843. MR Zbl

Received: 2018-12-05 Revised: 2019-03-18 Accepted: 2019-03-21

keoulaur@gvsu.edu *Department of Mathematics, Grand Valley State University, Allendale, MI, United States*

shaned@mail.gvsu.edu *Michigan State University, East Lansing, MI, United States*

INVOLVE 12:7 (2019)
dx.doi.org/10.2140/involve.2019.12.1183

On some obstructions of flag vector pairs (f_1, f_{04}) of 5-polytopes

Hye Bin Cho and Jin Hong Kim

(Communicated by Joshua Cooper)

Motivated by the recent work of Sjöberg and Ziegler, who obtained a complete characterization of the pairs (f_0, f_{03}) of flag numbers for 4-polytopes, in this paper we give some new results about the possible flag vector pairs (f_1, f_{04}) of 5-polytopes.

1. Introduction

Let P be a d-dimensional convex polytope. For each $0 \le i \le d-1$, let $f_i(P)$ denote the number of i-dimensional faces of P. One fundamental combinatorial invariant of P is the *f-vector*

$$f(P) = (f_0(P), f_1(P), \ldots, f_{d-1}(P)),$$

and characterizing all possible f-vectors of convex polytopes has been one of the central problems in convex geometry. For simplicity, throughout the paper a d-dimensional convex polytope will be called a *d-polytope*.

Let $\mathcal{F}^d$ denote the set of all f-vectors of d-polytopes, and let $\Pi_{i,j}(\mathcal{F}^d)$ denote the projection of $\mathcal{F}^d$ onto the coordinates f_i and f_j. Steinitz [1906] completely determined all possible f-vectors of 3-polytopes:

Theorem 1.1. *The set $\Pi_{0,1}(\mathcal{F}^3)$ of all f-vectors (f_0, f_1) of 3-polytopes is equal to*

$$\left\{(v, e) \;\middle|\; \tfrac{3}{2}v \le e \le 3v - 6\right\}.$$

In dimensions $d \ge 4$, any d-polytope P satisfies

$$\frac{d}{2} f_0(P) \le f_1(P) \le \binom{f_0(P)}{2}. \tag{1-1}$$

However, any complete determination of all possible f-vectors of d-polytopes for $d \ge 4$ is still elusive. As some partial results, for $d = 4$ the projections of the f-vector onto two of the four coordinates have been determined by Grünbaum

MSC2010: 52B05, 52B11.
Keywords: polytopes, f-vectors, flag vectors, flag vector pairs, stacking, truncating.

[1967], Barnette and Reay [1973], and Barnette [1974] (see [Sjöberg and Ziegler 2018, Section 2] for more details).

Kusunoki and Murai [2019] characterized the first two entries of the f-vectors of 5-polytopes.

Theorem 1.2. *Let* $L = \left\{\left(v, \left[\frac{5}{2}v + 1\right]\right) \mid v \geq 7\right\}$, *and let*

$$G = \{(8, 20), (9, 25), (13, 35)\}.$$

Then we have

$$\Pi_{0,1}(\mathcal{F}^5) = \left\{(v, e) \;\middle|\; \frac{5}{2}v \leq e \leq \binom{v}{2}\right\} \backslash (L \cup G).$$

The same result has been independently proved by Pineda-Villavicencio, Ugon, and Yost [2018] (see also [Pineda-Villavicencio, Ugon, and Yost 2019]).

For a subset S of $\{0, 1, 2, \ldots, d-1\}$, let $f_S(P)$ denote the number of chains

$$F_1 \subset F_2 \subset \cdots \subset F_r$$

of faces F_i, $1 \leq i \leq r$, of P such that

$$S = \{\dim F_1, \dim F_2, \ldots, \dim F_r\}.$$

The *flag vector* of P is defined to be

$$(f_S(P))_{S \subset \{0,1,2,\ldots,d-1\}}.$$

For the sake of simplicity, from now on we use the notation $f_{i_1 i_2 \ldots i_k}(P)$ instead of $f_{\{i_1, i_2, \ldots, i_k\}}(P)$ for any subset $\{i_1, i_2, \ldots, i_k\}$ of $\{0, 1, 2, \ldots, d-1\}$.

In this paper, for any two subsets S_1 and S_2 of $\{0, 1, 2, \ldots, d-1\}$ a pair $(f_{S_1}(P), f_{S_2}(P))$, or simply (f_{S_1}, f_{S_2}), of flag numbers of P will be called a *flag vector pair*. More generally, for any k not necessarily mutually disjoint subsets $S_1, S_2, \ldots, S_k$ of $\{0, 1, 2, \ldots, d-1\}$, a k-tuple

$$(f_{S_1}(P), f_{S_2}(P), \ldots, f_{S_k}(P)),$$

or simply $(f_{S_1}, f_{S_2}, \ldots, f_{S_k})$, of flag numbers of P will be called a *flag vector k-tuple*.

We denote by $\Pi_{S_1, S_2, \ldots, S_k}$ the projection of the flag vector $(f_S(P))_{S \subset \{0,1,2,\ldots,d-1\}}$ onto its coordinates $f_{S_1}, f_{S_2}, \ldots, f_{S_k}$. We call $(f_{S_1}, f_{S_2}, \ldots, f_{S_k})$ a *polytopal flag vector k-tuple* if

$$(f_{S_1}, f_{S_2}, \ldots, f_{S_k})$$

belongs to the image of the set of all flag vectors of d-dimensional polytopes under the projection map $\Pi_{S_1, S_2, \ldots, S_k}$, that is, if there is a d-polytope P such that

$$(f_{S_1}(P), f_{S_2}(P), \ldots, f_{S_k}(P)) = (f_{S_1}, f_{S_2}, \ldots, f_{S_k}).$$

Recently, Sjöberg and Ziegler [2018] obtained a complete characterization of the pairs (f_0, f_{03}) of flag numbers for 4-polytopes:

Theorem 1.3. *Let*

$$E = \left\{ \begin{matrix} (6,24),(6,25),(6,28),(7,28),(7,30),(7,31),(7,33),(7,34),(7,37), \\ (7,40),(8,33),(8,34),(8,37),(8,40),(9,37),(9,40),(10,40),(10,43) \end{matrix} \right\}.$$

Then the set of all flag vector pairs (f_0, f_{03}) *of* 4*-polytopes is equal to*

$$\left\{ (f_0, f_{03}) \;\middle|\; \begin{matrix} 20 \le 4f_0 \le f_{03} \le 2f_0(f_0-3), \\ f_{03} \ne 2f_0(f_0-3)-k,\ k \in \{1,2,3,5,6,9,13\} \end{matrix} \right\} \backslash E.$$

For the proof of Theorem 1.3, the classification of all combinatorial types of 4-polytopes with up to eight vertices by Altshuler and Steinberg [1984; 1985] played an important role.

Our primary aim of this paper is to provide some new results about the flag vector pairs (f_1, f_{04}) of 5-polytopes:

Theorem 1.4. *Let* P *be a* 5*-polytope. Then the flag vector pairs* (f_1, f_{04}) *of* 5*-polytopes satisfy the following inequalities*:

(1) *For a given flag number* $f_{04}(P)$, *we have*

$$\tfrac{5}{4}\left(7+\sqrt{1+\tfrac{4}{5}f_{04}(P)}\right) \le f_1(P) < \tfrac{1}{4}f_{04}(P)(f_{04}(P)-3). \tag{1-2}$$

(2) *For a given flag number* $f_1(P)$, *we have*

$$\tfrac{1}{2}\left(3+\sqrt{9+16f_1(P)}\right) < f_{04}(P) \le \tfrac{4}{5}f_1(P)^2 - 14f_1(P) + 60. \tag{1-3}$$

Remark 1.5. (1) The lower (resp. upper) bound of the flag vector pairs (f_1, f_{04}) given in Theorem 1.4(1) (resp. (2)) are very sharp, since there is an explicit example, such as a 5-simplex with $(f_1, f_{04}) = (15, 30)$, which satisfies the equalities in (1-2) and (1-3).

(2) The upper (resp. lower) bound of the flag vector pairs (f_1, f_{04}) given in Theorem 1.4(1) (resp. (2)) might be improved further by using much sharper inequality instead of $\sum_{i=1}^{k} x_i^2 < \left(\sum_{i=1}^{k} x_i\right)^2$ for any positive $x_i > 0$ with $1 \le i \le k$ or by any other means (see Lemma 2.1 for more details). In this paper, we do not pursue this issue further, though.

(3) The question of whether or not all vector pairs (f_1, f_{04}) satisfying the inequalities (1-2) and (1-3) given in Theorem 1.4 are flag vector pairs of 5-polytopes is unknown, and the technique of this paper is insufficient to answer such a question.

This paper is organized as follows. In Section 2, we give a proof of Theorem 1.4 by a series of lemmas. In Section 3, we provide some concrete examples of 5-polytopes satisfying the inequalities given in Theorem 1.4 for the flag vector pairs (f_1, f_{04}) of 5-polytopes. In order to construct such examples, we make use of the well-known stacking and truncating operations.

2. Proof of Theorem 1.4

We begin with the following lemmas.

Lemma 2.1. *The flag vector pair* $(f_1(P), f_{04}(P))$ *of a* 5*-polytope* P *satisfies*

$$f_1(P) < \tfrac{1}{4} f_{04}(P)(f_{04}(P) - 3).$$

Proof. Let F be any facet of the 5-polytope P. Then it follows from [Sjöberg and Ziegler 2018, Theorem 2.1] that

$$f_3(F) \le \tfrac{1}{2} f_0(F)(f_0(F) - 3).$$

Thus it is easy to obtain

$$\sum_{F \subset P} f_3(F) \le \frac{1}{2} \sum_{F \subset P} f_0^2(F) - \frac{3}{2} \sum_{F \subset P} f_0(F). \tag{2-1}$$

Since

$$\sum_{i=1}^{k} x_i^2 < \left(\sum_{i=1}^{k} x_i \right)^2$$

for any positive x_i $(1 \le i \le k)$, it follows from (2-1) that

$$\begin{aligned} f_{34}(P) = \sum_{F \subset P} f_3(F) &< \frac{1}{2} \left(\sum_{F \subset P} f_0(F) \right)^2 - \frac{3}{2} \sum_{F \subset P} f_0(F) \\ &= \tfrac{1}{2} f_{04}(P)^2 - \tfrac{3}{2} f_{04}(P). \end{aligned} \tag{2-2}$$

By considering the dual polytope P^* of P, by (2-2) we can obtain

$$2 f_1(P^*) = f_{01}(P^*) < \tfrac{1}{2} f_{04}(P^*)(f_{04}(P^*) - 3).$$

Since P is an arbitrary polytope, so is its dual P^*. Therefore, we can obtain

$$f_1(P) < \tfrac{1}{4} f_{04}(P)(f_{04}(P) - 3). \qquad \square$$

Lemma 2.2. *The flag vector pair* $(f_0(P), f_{04}(P))$ *of a* 5*-polytope* P *satisfies*

$$5 f_0(P) \le f_{04}(P) \le 5(f_0(P) - 3)(f_0(P) - 4).$$

Proof. Note first that every vertex of a d-polytope meets at least d facets. Thus we have $5 f_0(P) \le f_{04}(P)$, where equality holds if and only if P is a simple polytope.

On the other hand, it follows from [Sjöberg and Ziegler 2018, Lemma 2.6] (or [Billera and Björner 1997, Theorem 18.5.9]) that for any d-polytope Q with n vertices and for any subset $S \subset \{0, 1, 2, \ldots, d-1\}$ we have

$$f_S(Q) \le f_S(C_d(n)),$$

where $C_d(n)$ denotes the d-dimensional cyclic polytope with $n = f_0(Q)$ vertices. Hence, we have

$$f_{04}(P) \le f_{04}(C_5(n)) = 5f_4(C_5(n)). \tag{2-3}$$

Here, the second equality holds because $C_5(n)$ is simplicial, and the first inequality becomes an equality if and only if P is neighborly.

On the other hand, by using the formula in [Buchstaber and Panov 2002, Lemma 1.34] we can directly calculate

$$\begin{aligned} f_4(C_5(n)) &= \sum_{q=0}^{2} \binom{q}{0}\binom{n+q-6}{q} + \sum_{p=0}^{2} \binom{5-p}{5-p}\binom{n+p-6}{p} \\ &= (n-3)(n-4). \end{aligned}$$

Hence, it follows from (2-3) that

$$f_{04}(P) \le 5f_4(C_5(n)) = 5(f_0(P)-3)(f_0(P)-4). \qquad \square$$

Lemma 2.3. *The flag vector pair $(f_1(P), f_{04}(P))$ of a 5-polytope P satisfies*

$$f_1(P) \ge \tfrac{5}{4}\big(7 + \sqrt{1 + \tfrac{4}{5}f_{04}(P)}\big).$$

Proof. By Lemma 2.2, we have

$$f_0(P)^2 - 7f_0(P) + 12 - \tfrac{1}{5}f_{04}(P) \ge 0.$$

Thus, since $f_0(P) \ge 6$, it is easy to obtain

$$f_0(P) \ge \tfrac{1}{2}\big(7 + \sqrt{1 + \tfrac{4}{5}f_{04}(P)}\big). \tag{2-4}$$

Recall now that $f_0(P) \le \frac{2}{5}f_1(P)$ by (1-1). Hence, it follows from (2-4) that

$$f_1(P) \ge \tfrac{5}{4}\big(7 + \sqrt{1 + \tfrac{4}{5}f_{04}(P)}\big),$$

as desired. $\square$

Theorem 1.4(1) is an immediate consequence of Lemmas 2.1 and 2.3.

Next, we want to prove Theorem 1.4(2). We begin with the *generalized Dehn–Sommerville equations*, given in the following theorem (see [Sjöberg and Ziegler 2018, Theorem 2.4] and [Bayer and Billera 1985, Theorem 2.1] for more details).

Theorem 2.4. *Let P be a d-polytope, and let S be a subset of $\{0, 1, 2, \ldots, d-1\}$. If $\{i, k\}$ is a subset of $S \cup \{-1, d\}$ such that $i < k-1$ and such that there is no $j \in S$ for which $i < j < k$, then*

$$\sum_{j=i+1}^{k-1} (-1)^{j-i-1} f_{S\cup\{j\}}(P) = f_S(P)(1 - (-1)^{k-i-1}).$$

Corollary 2.5. *The flag vector* 4*-tuple* $(f_{01}(P), f_{02}(P), f_{03}(P), f_{04}(P))$ *of a* 5*-polytope* P *satisfies*

$$f_{01}(P) - f_{02}(P) + f_{03}(P) - f_{04}(P) = 0. \tag{2-5}$$

Proof. Let $S = \{0\}$, $i = 0$, and $k = 5$. By applying Theorem 2.4 to these choices of S, i, and k, it is immediate to obtain (2-5). □

Lemma 2.6. *The flag vector* 3*-tuple* $(f_1(P), f_{02}(P), f_{04}(P))$ *of a* 5*-polytope* P *satisfies*

$$2f_1(P) - f_{02}(P) + f_{04}(P) \le 0.$$

Proof. As in the proof of Lemma 2.1, let F denote any facet of P. By [Sjöberg and Ziegler 2018, Theorem 2.2], we have

$$f_1(F) \ge 2f_0(F).$$

Thus, it is easy to obtain

$$f_{14}(P) = \sum_{F \subset P} f_1(F) \ge 2 \sum_{F \subset P} f_0(F) = 2f_{04}(P). \tag{2-6}$$

By duality, it follows from (2-6) that

$$f_{03}(P) \ge 2f_{04}(P). \tag{2-7}$$

On the other hand, by Corollary 2.5 together with (2-6) we also have

$$\begin{aligned} f_{04}(P) &= f_{01}(P) - f_{02}(P) + f_{03}(P) \\ &\ge f_{01}(P) - f_{02}(P) + 2f_{04}(P). \end{aligned}$$

Since $2f_1(P) = f_{01}(P)$, finally we obtain

$$2f_1(P) - f_{02}(P) + f_{04}(P) \le 0,$$

as desired. □

Lemma 2.7. *The flag vector pair* $(f_0(P), f_{02}(P))$ *of a* 5*-polytope* P *satisfies*

$$f_{02}(P) \le 6(f_0(P)^2 - 6f_0(P) + 10).$$

Proof. As in the proof of Lemma 2.2, by applying the upper bound theorem stated in [Sjöberg and Ziegler 2018, Lemma 2.6] (see also [Billera and Björner 1997, Theorem 18.5.9]) we obtain

$$f_{02}(P) \le f_{02}(C_5(n)) = 3f_2(C_5(n)),$$

where $f_0(P) = n$ and the fact that $C_5(n)$ is a simplicial polytope was used in the last equality.

On the other hand, by using the formula of $f_2(C_5(n))$ given in [Buchstaber and Panov 2002, Lemma 1.34] it is straightforward to compute

$$\begin{aligned} f_2(C_5(n)) &= \sum_{q=0}^{2}\binom{q}{2}\binom{n+q-6}{q} + \sum_{p=0}^{2}\binom{5-p}{2}\binom{n+p-6}{p} \\ &= \binom{n-4}{2} + \binom{5}{2}\binom{n-6}{0} + \binom{4}{2}\binom{n-5}{1} + \binom{3}{2}\binom{n-4}{2} \\ &= 2(n^2-6n+10) = 2(f_0(P)^2 - 6f_0(P) + 10). \end{aligned}$$

□

Lemma 2.8. *The flag vector pair $(f_1(P), f_{04}(P))$ of a 5-polytope P satisfies*

$$f_{04}(P) \le \tfrac{1}{25}(24f_1(P)^2 - 410f_1(P) + 1500).$$

Proof. By Lemma 2.6, it is easy to obtain

$$\begin{aligned} f_{04}(P) &\le -2f_1(P) + f_{02}(P) \\ &\le -2f_1(P) + 6(f_0(P)^2 - 6f_0(P) + 10) \\ &\le -2f_1(P) + 6\big(\tfrac{4}{25}f_1(P)^2 - \tfrac{12}{5}f_1(P) + 10\big) \\ &= \tfrac{1}{25}(24f_1(P)^2 - 410f_1(P) + 1500), \end{aligned}$$

where we used $f_0(P) \le \frac{2}{5}f_1(P)$ and $f_0(P) \ge 6$ in the third inequality. □

In fact, it turns out that for any values of $f_1(P) > 15$ the upper bound of $f_{04}(P)$ given in Lemma 2.8 can be improved further by using (1-2).

Lemma 2.9. *The flag vector pair $(f_1(P), f_{04}(P))$ of a 5-polytope satisfies*

$$f_{04}(P) \le \tfrac{4}{5}f_1(P)^2 - 14f_1(P) + 60.$$

Proof. For the proof, note that by Lemma 2.3 we have

$$f_1(P) \ge \tfrac{5}{4}\big(7 + \sqrt{1 + \tfrac{4}{5}f_{04}(P)}\big).$$

Thus, it is easy to obtain

$$f_{04}(P) \le \tfrac{4}{5}f_1(P)^2 - 14f_1(P) + 60.$$

□

For any 5-polytopes, $f_1(P) \ge 15$. Thus it is straightforward to show that

$$\tfrac{4}{5}f_1(P)^2 - 14f_1(P) + 60 \le \tfrac{1}{25}(24f_1(P)^2 - 410f_1(P) + 1500),$$

where equality holds if and only if $f_1(P) = 15$.

Finally, we are in a position to give a proof of Theorem 1.4(2):

Theorem 2.10. *Given a flag number $f_1(P)$ of a 5-polytope P, $f_{04}(P)$ satisfies*

$$\tfrac{1}{2}(3 + \sqrt{9 + 16f_1(P)}) < f_{04}(P) \le \tfrac{4}{5}f_1(P)^2 - 14f_1(P) + 60.$$

Proof. By Lemma 2.9, it suffices to prove the first inequality. Indeed, recall from Lemma 2.1 that we have

$$4f_1(P) < f_{04}(P)(f_{04}(P)-3), \quad \text{i.e., } f_{04}(P)^2 - 3f_{04}(P) - 4f_1(P) > 0.$$

This immediately implies

$$f_{04}(P) > \tfrac{1}{2}(3+\sqrt{9+16f_1(P)}).$$ □

3. Some examples

The aim of this section is to provide some examples of 5-polytopes whose flag vector pairs (f_1, f_{04}) satisfy the inequalities (1-2) and (1-3) given in Theorem 1.4. In order to construct such examples, we use the well-known operations of stacking and truncating. In many instances, these operations turn out to be essential in finding new examples of polytopes for possible polytopal flag vector pairs.

To begin with, we have the following lemma.

Lemma 3.1. *Let P be a 5-polytope with at least one simple facet F, and let v be a point beyond F and beneath all other facets of P. Let Q be the 5-polytope obtained by stacking the vertex v over P; i.e., let Q be the convex hull of v and P. Then we have the identities*

$$f_0(Q) = f_0(P) + 1,$$
$$f_1(Q) = f_1(P) + 5,$$
$$f_{04}(Q) = f_{04}(P) + 20.$$

Proof. By the way of the construction of Q, it suffices to show the last identity. To see it, note first that F is a 4-simplex with five vertices. If we apply the stacking operation to P with such a vertex v over F, then it is easy to see that the flag number f_{04} increases by $5\binom{5}{4}$ and decreases by 5. Thus the net change of f_{04} is equal to 20, and so we have

$$f_{04}(Q) = f_{04}(P) + 20.$$ □

Let P be a d-polytope with a vertex v, and let H be a hyperplane intersecting the interior of P such that on one side of H the only vertex of P is v. Then we can obtain a new polytope Q by cutting off the side of H that contains v. This operation of obtaining a new polytope is called a *truncating at a vertex*.

The following lemma holds.

Lemma 3.2. *Let P be a 5-polytope with at least one simple vertex v, and let R be the 5-polytope obtained by truncating the vertex v from P. Then we have the identities*

$$f_0(R) = f_0(P) + 4,$$
$$f_1(R) = f_1(P) + 10,$$
$$f_{04}(R) = f_{04}(P) + 20.$$

Proof. By the way of the construction of R, once again it suffices to prove the last equality. Note first that by the truncating operation we have five new vertices, all of which are simple. Thus the flag number f_{04} increases by 5×5 and decreases by 5 coming from the old vertex v. This implies $f_{04}(R) = f_{04}(P) + 20$, as required. □

Note that the polytopes obtained through stacking over a simple vertex v and truncating at v all have a simple vertex and a simplex facet. Thus we can repeatedly stack vertices on simplex facets and truncate simple vertices.

With these understood, let P be a 5-polytope P with a 4-simplex facet and a simple vertex. By truncating simple vertices l times and stacking vertices on 4-simplex facets k times inductively, we can obtain a new 5-polytope Q with the flag vector pair

$$(f_1(Q), f_{04}(Q)) = (f_1(P) + 5k + 10l, f_{04}(P) + 20k + 20l), \quad k, l \geq 0. \quad (3\text{-}1)$$

Let $n = k + l$. Then it follows from (3-1) that

$$(f_1(Q), f_{04}(Q)) = (f_1(P) + 10n - 5k, f_{04}(P) + 20n), \quad n \geq 0,\ 0 \leq k \leq n. \quad (3\text{-}2)$$

As a special case, let P be a 5-simplex. Then the flag vector pair $(f_1(P), f_{04}(P))$ is equal to $(15, 30)$. Thus, by (3-2) we can obtain the flag vector pair

$$(f_1(Q), f_{04}(Q)) = (10n - 5k + 15, 20n + 30), \quad n \geq 0,\ 0 \leq k \leq n.$$

One may check directly that the flag vector pair $(f_1(Q), f_{04}(Q))$ satisfies the inequalities (1-2) and (1-3) given in Theorem 1.4.

Acknowledgements

The authors are very grateful to the anonymous referee for valuable comments on this paper. This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2016R1D1A1B03930639).

References

[Altshuler and Steinberg 1984] A. Altshuler and L. Steinberg, "Enumeration of the quasisimplicial 3-spheres and 4-polytopes with eight vertices", *Pacific J. Math.* **113**:2 (1984), 269–288. MR Zbl

[Altshuler and Steinberg 1985] A. Altshuler and L. Steinberg, "The complete enumeration of the 4-polytopes and 3-spheres with eight vertices", *Pacific J. Math.* **117**:1 (1985), 1–16. MR Zbl

[Barnette 1974] D. Barnette, "The projection of the f-vectors of 4-polytopes onto the (E, S)-plane", *Discrete Math.* **10** (1974), 201–216. MR Zbl

[Barnette and Reay 1973] D. Barnette and J. R. Reay, "Projections of f-vectors of four-polytopes", *J. Combinatorial Theory Ser. A* **15** (1973), 200–209. MR Zbl

[Bayer and Billera 1985] M. M. Bayer and L. J. Billera, "Generalized Dehn–Sommerville relations for polytopes, spheres and Eulerian partially ordered sets", *Invent. Math.* **79**:1 (1985), 143–157. MR Zbl

[Billera and Björner 1997] L. J. Billera and A. Björner, "Face numbers of polytopes and complexes", pp. 291–310 in *Handbook of discrete and computational geometry*, edited by J. E. Goodman and J. O'Rourke, CRC, Boca Raton, FL, 1997. MR Zbl

[Buchstaber and Panov 2002] V. M. Buchstaber and T. E. Panov, *Torus actions and their applications in topology and combinatorics*, Univ. Lecture Series **24**, Amer. Math. Soc., Providence, RI, 2002. MR Zbl

[Grünbaum 1967] B. Grünbaum, *Convex polytopes*, Pure Appl. Math. **16**, Interscience, New York, 1967. MR Zbl

[Kusunoki and Murai 2019] T. Kusunoki and S. Murai, "The numbers of edges of 5-polytopes with a given number of vertices", *Ann. Comb.* **23**:1 (2019), 89–101. MR Zbl

[Pineda-Villavicencio, Ugon, and Yost 2018] G. Pineda-Villavicencio, J. Ugon, and D. Yost, "The excess degree of a polytope", *SIAM J. Discrete Math.* **32**:3 (2018), 2011–2046. MR Zbl

[Pineda-Villavicencio, Ugon, and Yost 2019] G. Pineda-Villavicencio, J. Ugon, and D. Yost, "Lower bound theorems for general polytopes", *European J. Combin.* **79** (2019), 27–45. MR Zbl

[Sjöberg and Ziegler 2018] H. Sjöberg and G. M. Ziegler, "Characterizing face and flag vector pairs for polytopes", *Discrete Comput. Geom.* (online publication November 2018).

[Steinitz 1906] E. Steinitz, "Über die Eulerschen Polyederrelationen", *Arch. Math. Phys.* **11** (1906), 86–88. Zbl

Received: 2018-12-19 Revised: 2019-06-18 Accepted: 2019-06-22

jinhkim11@gmail.com *Department of Mathematics Education, Chosun University, Gwangju, South Korea*

gpqls010@daum.net *Department of Mathematics Education, Chosun University, Gwangju, South Korea*

INVOLVE 12:7 (2019)
dx.doi.org/10.2140/involve.2019.12.1193

Benford's law beyond independence: tracking Benford behavior in copula models

Rebecca F. Durst and Steven J. Miller

(Communicated by Stephan Garcia)

Benford's law describes a common phenomenon among many naturally occurring data sets and distributions in which the leading digits of the data are distributed with the probability of a first digit of d base B being $\log_B((d+1)/d)$. As it often successfully detects fraud in medical trials, voting, science and finance, significant effort has been made to understand when and how distributions exhibit Benford behavior. Most of the previous work has been restricted to cases of independent variables, and little is known about situations involving dependence. We use copulas to investigate the Benford behavior of the product of n dependent random variables. We develop a method for approximating the Benford behavior of a product of n dependent random variables modeled by a copula distribution C and quantify and bound a copula distribution's distance from Benford behavior. We then investigate the Benford behavior of various copulas under varying dependence parameters and number of marginals. Our investigations show that the convergence to Benford behavior seen with independent random variables as the number of variables in the product increases is not necessarily preserved when the variables are dependent and modeled by a copula. Furthermore, there is strong indication that the preservation of Benford behavior of the product of dependent random variables may be linked more to the structure of the copula than to the Benford behavior of the marginal distributions.

1. Introduction

Benford's law of digit bias applies to many commonly encountered data sets and distributions. A set of data $\{x_i\}_{i\in I}$ is said to be *Benford base* B if the probability of observing a value x_i in the set with the first digit d (where d is any integer from 1 to $B-1$) is given by the equation

$$\text{Prob(first digit of } \{x_i\}_{i\in I} \text{ is } d) \text{ base } B = \log_B\Big(\frac{d+1}{d}\Big). \tag{1-1}$$

MSC2010: 11K99, 60E99.
Keywords: Benford's law, probability, theoretical statistics .

These probabilities monotonically decrease; e.g., in base 10 there is a leading digit of 1 about 30.103% of the time and a leading digit of 9 about 4.576% of the time.

Benford's law was discovered in 1881 by the astronomer-mathematician Simon Newcomb who, looking at his logarithm table, observed earlier pages were more heavily worn than later pages. As logarithm tables are organized by leading digit, this led him to conclude that values with leading digit 1 occurred more commonly than values with higher leading digits. These observations were mostly forgotten for fifty years, when Benford [1938] published his work detailing similar biases in a variety of settings. Since then, the number of fields where Benford behavior is seen has rapidly grown, including accounting, biology, computer science, economics, mathematics, physics and psychology to name a few; see [Benford 2009; Berger and Hill 2015; Miller 2015; Nigrini 1999; Raimi 1976] for a development of the general theory and many applications. This prevalence of Benford's law, particularly in naturally occurring data sets and common distributions, has allowed it to become a useful tool in detecting fraud. One notable example of this was its use in 2009 to find evidence suggesting the presence of fraud in the Iranian elections [Battersby 2009]. While Benford's law cannot prove that fraud happened, it is a useful tool for determining which sets of data are suspicious enough to merit further investigation (which is of great importance given finite resources); see for example [Nigrini and Mittermaier 1997; Singleton 2011].

To date, most of the work on the subject has involved independent random variables or deterministic processes (see though [Becker et al. 2018; Iafrate et al. 2015] for work on dependencies in partition problems). Our goal below is to explore dependent random variables through copulas, quantifying the connections between various relations and Benford behavior.

Copulas are multivariate probability distributions restricted to the unit hypercube by transforming the marginals into uniform random variables via the probability integral transform (see Section 2 for precise statements). The term copulas was first defined by Abe Sklar in 1959, when he published what is now known as Sklar's theorem (see Theorem 2.7), though similar objects were present in the work of Wassily Hoeffding as early as 1940. Sklar described their purpose as linking n-dimensional distributions with their one-dimensional margins. See [Nelsen 2006] for a detailed account of the presence and evolution of copulas.

Fisher [1997] writes, "Copulas [are] of interest to statisticians for two main reasons: Firstly, as a way of studying scale-free measures of dependence; and secondly, as a starting point for constructing families of bivariate distributions, sometimes with a view to simulation." More specifically, copulas are widely used in application in fields such as economics and actuarial studies; for example, [Kpanzou 2007] describes applications in survival analysis and extreme value theory, and [Wu et al. 2007] details the use of Archimedean copulas in economic modeling

and risk management. Thus, as copulas are a convenient and useful way to model dependent random variables, they are often employed in fields relating to finance and economics. Since many of these areas are also highly susceptible to fraud, it is worth exploring connections between copulas and Benford's law, with the goal to develop data integrity tests.

Essentially, since so many dependencies may be modeled through copulas, it is natural to ask when and how often these structures will display Benford behavior. In this paper, we investigate when data modeled by a copula is close to Benford's law by developing a method for approximating Benford behavior. In Section 3, we develop this method for the product of n random variables whose joint distribution is modeled by the copula C. We then apply this method in Section 4 to directly investigate Benford behavior for various copulas and dependence parameters. We conclude that Benford behavior depends heavily on the structure of the copula. We use goodness of fit measures to show both numerically and graphically that the product of many random variables with dependence modeled by a copula will not necessarily level-off like products of independent random variables, the log of which we may expect to become more uniform as the number of variables increases. The results of this paper extend current techniques for testing Benford's law to situations where independence is not guaranteed, allowing analyses like that carried out in [Cuff et al. 2015] on the Weibull distribution and in [Durst et al. 2016] on the inverse gamma distribution to be conducted in the case of n dependent random variables. In Section 5, we restrict ourselves to n-tuples of random variables in which at least one is a Benford distribution and develop a concept of distance between our joint distribution and a Benford distribution, thus developing a concept of distance from a Benford distribution in order to understand how much deviation from Benford one might expect of a particular distribution. We then provide an upper bound for this distance using the L^1 norm of the function

$$N(u_1, u_2, \ldots, u_n) = 1 - \frac{\partial^n C(u_1, u_2, \ldots, u_n)}{\partial u_1 \partial u_2 \ldots \partial u_n}.$$

In doing so, we draw an interesting connection between the distance from a Benford distribution and a copula's distance from the space of copulas for which $C_{uv}(u, v) = 1$ for all u, v in $[0, 1]$.

2. Terms and definitions

We abbreviate *probability density function* by PDF and *cumulative distribution function* by CDF, and assume all CDFs are uniformly or absolutely continuous. All results below are standard; see the references for proofs.

General mathematics and Benford's law.

Lemma 2.1 (Barbalat's lemma [Fontes and Magni 2004, Lemma 2.1]). *Let $t \mapsto F(t)$ be a differentiable function with a finite limit as $t \to \infty$. If F' is uniformly continuous, then $F'(t) \to 0$ as $t \to \infty$.*

Definition 2.2 (scientific notation). Any real number, x, can be written in the form

$$x = S_B(x) \cdot B^n, \tag{2-1}$$

where n is an integer and $S_B(x) < 10$. We call B the *base* and $S_B(x)$ the *significand*.

We define strong Benford's law base B; see, for example, [Berger and Hill 2015; Miller 2015]. This is the definition we primarily use in Section 3; strong indicates that we are studying the entire significand of the number and not just its first digit. In Section 5, we will provide insight into how one may define a weaker version of Benford's law that permits the probabilities to be within ϵ of the theoretical Benford probabilities.

Definition 2.3 (strong Benford's law [Miller 2015, Definition 1.6.1]). A data set satisfies the strong Benford's law base B if the probability of observing a leading digit of at most s in base B is $\log_B s$.

Theorem 2.4 (absorptive property of Benford's law [Tao 2010, page 56]). *Let X and Y be **independent** random variables. If X obeys Benford's law, then the product $W = XY$ obeys Benford's law regardless of whether or not Y obeys Benford's law.*

Copulas. All theorems and definitions in this section are from [Nelsen 2006] unless otherwise stated.

Remark 2.5. In [Nelsen 2006], functions are defined on the *extended real line*, $[-\infty, \infty]$; thus $f(t)$ is defined when $t = \pm\infty$. We use this notation in order to maintain consistency with that work, as it is one of the central texts in copula theory.

Definition 2.6 (n-dimensional copula). An *n-dimensional copula*, C, is a function satisfying the following properties:

(1) The domain of C is $[0, 1]^n$.

(2) (n-increasing) The n-th order difference of C is greater than or equal to zero.

(3) (grounded) $C(u_1, u_2, \ldots, u_n) = 0$ if $u_k = 0$ for at least one k in $\{1, 2, \ldots, n\}$.

(4) $C(1, 1, \ldots, 1, u_k, 1, , \ldots, 1) = u_k$ for some k in $\{1, 2, \ldots, n\}$.

Theorem 2.7 (Sklar's theorem [Nelsen 2006, Theorem 2.10.9]). *Let H be an n-dimensional distribution function with marginal CDFs $F_1, F_2, \ldots, F_n$. Then there exists an n-copula C such that for all $(x_1, x_2, \ldots, x_n)$ in $[-\infty, \infty]^n$,*

$$H(x_1, x_2, \ldots, x_n) = C(F_1(x_1), F_2(x_2), \ldots, F_n(x_n)). \tag{2-2}$$

If all F_i are continuous, then C is unique; otherwise, C is uniquely determined on $\text{Range}(F_1) \times \text{Range}(F_2) \times \cdots \times \text{Range}(F_n)$. *Conversely, if C is a copula and $F_1, F_2, \ldots, F_n$ are cumulative distribution functions, then the function H defined by (2-2) is a distribution function with marginal cumulative distribution functions $F_1, F_2, \ldots, F_n$.*

Theorem 2.8 (extension of [Nelsen 2006, Theorem 2.4.2]). *Let $X_1, X_2, \ldots, X_n$ be continuous random variables. Then they are independent if and only if their copula, $C_{X_1,X_2,\ldots,X_n}$, is given by $C_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) = \Pi(x_1, x_2, \ldots, x_n) = x_1 x_2 \cdots x_n$, where Π is called the* ***product copula****.*

Theorem 2.9 (extension of [Nelsen 2006, Theorem 2.4.3]). *Let $X_1, X_2, \ldots, X_n$ be continuous random variables with copula $C_{X_1,X_2,\ldots,X_n}$. If $a_1, a_2, \ldots, a_n$ are strictly increasing on* $\text{Range}(X_1), \text{Range}(X_2), \ldots, \text{Range}(X_n)$, *respectively, then $C_{a_1(X_1),a_2(X_2),\ldots,a_n(X_n)} = C_{X_1,X_2,\ldots,X_n}$. Thus $C_{X_1,X_2,\ldots,X_n}$ is invariant under strictly increasing transformations of $X_1, X_2, \ldots, X_n$.*

Remark 2.10. For the following three definitions, see page 116 of [Nelsen 2006] for the 2-copula formulas and page 151 for the n-copula extension.

Definition 2.11 (Clayton family of copulas). An (n-dimensional) copula in the *Clayton family* is given by the equation

$$C(u_1, u_2, \ldots, u_n) = \max\{(u_1^{-\alpha} + u_2^{-\alpha} + \cdots + u_n^{-\alpha} + n - 1)^{-1/\alpha}, 0\}, \tag{2-3}$$

where $\alpha \in [-1, \infty) \setminus \{0\}$ is a parameter related to dependence, with $\alpha = 0$ as the independence case.

Definition 2.12 (Ali–Mikhail–Haq family of copulas). An (n-dimensional) copula in the *Ali–Mikhail–Haq family* is given by the equation

$$C(u_1, u_2, \ldots, u_n) = \frac{(1-\alpha)}{\left(\prod_{i=1}^{n}(1-\alpha(1-u_i))/u_i\right) - \alpha}, \tag{2-4}$$

where $\alpha \in [-1, 1)$ is a parameter related to dependence, with $\alpha = 0$ as the independence case.

Definition 2.13 (Gumbel–Barnett family of copulas). An (n-dimensional) copula in the *Gumbel–Barnett family* is given by the equation

$$C(u_1, u_2, \ldots, u_n) = \exp\frac{1-(1-\alpha\log u_1)(1-\alpha\log u_2)\cdots(1-\alpha\log u_n)}{\alpha}, \tag{2-5}$$

where $\alpha \in (0, 1]$ is a parameter related to dependence, with $\alpha = 0$ as the independence case.

3. Testing for Benford behavior of a product

We state the results below in arbitrary dimensions but for notational convenience give the proofs for just two dimensions as the generalization is straightforward.

Let $X_1, \dots, X_n$ be continuous random variables with CDFs $F_{X_1}(x_1), \dots, F_{X_n}(x_n)$. Let their joint PDF be $H_{X_1,\dots,X_n}(X_1, \dots, X_n)$. By Theorem 2.7, we know there exists a copula C such that

$$H_{X_1,\dots,X_n}(X_1, \dots, X_n) = C(F_{X_1}(X_1), \dots, F_{X_n}(X_n)). \tag{3-1}$$

Assume $X_1, \dots, X_n$ are such that their copula C is *absolutely continuous*. This allows us to define the joint probability density function [Nelsen 2006, page 27] by $\partial^n C/(\partial x_1 \cdots \partial x_n)$. Furthermore, we restrict ourselves to X_i such that all F_{X_i} are uniformly continuous, as this allows us to use Lemma 2.1 to later ensure that the PDFs approach zero in their right- and left-end limits.

From here we have the following lemma.

Lemma 3.1. *Given $X_1, \dots, X_n$ positive, continuous random variables with joint distribution modeled by the absolutely continuous copula C, let $U_i = \log_B X_i$ for all $i \le n$ and for some base B, and let the CDFs of each U_i be $F_i(u_i)$. Also, let $f_i(u_i)$ be the PDF of U_i for all i. Finally, let*

$$\boldsymbol{u}_0 = (u_1, \dots, u_{n-1}, s + k - (u_1 + \cdots + u_{n-1})).$$

Then

$$\text{Prob}\left(\left(\sum_{i=1}^{n} U_i\right) \bmod 1 \le s\right)$$

$$= \int_0^s \sum_{k=-\infty}^{\infty} \int_{u_1=-\infty}^{\infty} \cdots \int_{u_{n-1}=-\infty}^{\infty} \left.\frac{\partial^n C(F_1(u_1), \dots, F_{n-1}(u_{n-1}), F_n(u_n))}{\partial u_1 \cdots \partial u_n}\right|_{\boldsymbol{u}_0} du_1 \cdots du_{n-1}.$$

Therefore, the PDF of $(U + V) \bmod 1$ is given by

$$\sum_{k=-\infty}^{\infty} \int_{u_1=-\infty}^{\infty} \cdots \int_{u_{n-1}=-\infty}^{\infty} \left.\frac{\partial^n C(F_1(u_1), \dots, F_{n-1}(u_{n-1}), F_n(u_n))}{\partial u_1 \cdots \partial u_n}\right|_{\boldsymbol{u}_0} du_1 \cdots du_{n-1}. \tag{3-2}$$

See Appendix 1 for the proof.

If (3-2) equals 1 for all s, then our product is Benford. If it is not identically equal to 1 for all s, then at each point we may assign a value ϵ_s that represents our distance from a Benford distribution. Thus we have

$$\epsilon_s = |1 - P|, \tag{3-3}$$

where P is the PDF given in (3-2). This formulation will form the basis of Section 5.

Unfortunately, the infinite sum and improper integral in (3-2) make it highly impractical to use in application unless we can determine a method to closely approximate them by a finite sum and finite integral. We note that (3-2) is a PDF, and so is $\partial^n C/(\partial x_1 \cdots \partial x_n)$, so we have the following properties (for notational convenience we state them in the two-dimensional case; similar results hold for n-dimensions).

(1) $$\int_0^1 \left(\sum_{k=-\infty}^{\infty} \int_{-\infty}^{\infty} C_{u_1u_2}(F_1(u_1), F_2(s+k-u_1)) f_1(u_1) f_2(s+k-u_1)\, du_1 \right) ds = 1.$$

(2) $$\sum_{k=-\infty}^{\infty} \int_{-\infty}^{\infty} C_{u_1u_2}(F_1(u_1), F_2(s+k-u_1)) f_1(u_1) f_2(s+k-u_1)\, du_1 \geq 0 \text{ for all } s.$$

(3) $$\int_{-\infty}^{\infty} C_{u_1u_2}(F_1(u_1), F_2(s+k-u_1)) f_1(u_1) f_2(s+k-u_1)\, du_1 \to 0 \text{ as } k \to \pm\infty.$$

(4) $$C_{u_1u_2}(F_1(u_1), F_2(s+k-u_1)) f_1(u_1) f_2(s+k-u_1) \to 0 \text{ as } u_1 \to \pm\infty.$$

Property (1) is simply the definition of a PDF, and property (2) is a direct result of the fact that a PDF is always positive. Properties (3) and (4) are required, under Lemma 2.1, by the convergence of the integral in property (1) and by the convergence of the sum.

From properties (3) and (4) and the definition of convergence we obtain the following.

Lemma 3.2 (approximating the PDF). *Given $U_1, \ldots, U_n$ continuous random variables modeled by the copula C with marginal CDFs $F_1, \ldots, F_n$ and PDFs $f_1, \ldots, f_n$, there exist $a_1, \ldots, a_{n-1}, b_1, \ldots, b_{n-1}$, and c_1 and c_2 completely dependent on the F_i such that $a_i < b_i$ for all i and $c_1 < c_2$ and*

$$\sum_{k=-\infty}^{\infty} \int_{u_1=-\infty}^{\infty} \cdots \int_{u_{n-1}=-\infty}^{\infty} \frac{\partial^n C(F_1(u_1), \ldots, F_{n-1}(u_{n-1}), F_n(u_n))}{\partial u_1 \cdots \partial u_n}\bigg|_{\boldsymbol{u}_0} du_1 \cdots du_{n-1}$$

$$= \sum_{k=c_1}^{c_2} \int_{u_1=a_1}^{b_1} \cdots \int_{u_{n-1}=a_{n-1}}^{b_{n-1}} \frac{\partial^n C(F_1(u_1), \ldots, F_{n-1}(u_{n-1}), F_n(u_n))}{\partial u_1 \cdots \partial u_n}\bigg|_{\boldsymbol{u}_0} du_1 \cdots du_{n-1} + E_{a,b,c}(s), \quad (3\text{-}4)$$

where $E_{a,b,c}(s) \to 0$ as each a_i and c_1 go to $-\infty$ and each b_i and c_2 go to ∞. Thus, for any $\epsilon > 0$, there exists (for each i) $|a_i|$, $|b_i|$, $|c_1|$, and $|c_2|$ large enough such that $|E_{a,b,c}(s)| \leq \epsilon$.

The proof of this claim can be found in Appendix 1.

The specific values of c_1, c_2, and each a_i and b_i are best determined by numerically testing the errors caused by truncating the interval using either known error bounds or appropriate software. As seen in Appendix 2, the values for these constants are often quite reasonable, as long as the functions decay fast enough in the limit.

Because s only ranges from 0 to 1, we can always find a value of s that maximizes $E_{a,b,c}$ for any given set of a, b, and c and set this to be the maximum error. Furthermore, since all f_i should have similar tail-end behavior, we do not have to worry about the divergence of one canceling out the divergence of the other. Thus, for this analysis to work, it is sufficient to understand the tail-end behavior of only one of the marginals.

In Appendix 2, we provide several examples of this method for testing for Benford behavior computationally with two variables.

4. Testing For Benford behavior: examples

Now that an effective method for testing the Benford behavior of copulas has been established, we investigate how this behavior varies for specific copulas and marginals. In all χ^2 tests, we follow standard procedure for multiple comparison problems. We are sampling our distribution at 12 values of s, necessitating 11 degrees of freedom, and we impose a significance level of 0.005, meaning we only accept a 0.5% probability of false rejection. Thus, we reject the hypothesis, specifically *we reject that the distribution displays Benford behavior*, if the χ^2-value exceeds 2.6. Our main interest, however, is to observe how and if these values trend towards this critical value.

Please note that the α used in this section is the dependence parameter of the copula and does not represent the significance level, as traditionally seen in statistical analysis.

2-copulas with varying dependence parameter. The following figures display the nonerror values of (3-4) at various values of s for three different copulas. The line in each plot indicates the constant function $y = 1$, which will be achieved if the product XY is exactly Benford. For each copula, we test three different pairings of marginals:

(A) $\log X \sim \mathrm{N}(0, 1)$ and $\log Y \sim \mathrm{Exp}(1)$.

(B) $\log X \sim \mathrm{Pareto}(1)$ and $\log Y \sim \mathrm{N}(0, 1)$.

(C) $\log X \sim \mathrm{Pareto}(1)$ and $\log Y \sim \mathrm{Exp}(1)$.

In each case, we vary the dependence parameter α and compare the results to the case of independence. Our Pareto distribution has scale parameter $x_m = 1$ and shape parameter $\alpha_p = 2$. We note that in some cases the axes must be adjusted to be able to show any change in the Benford behavior.

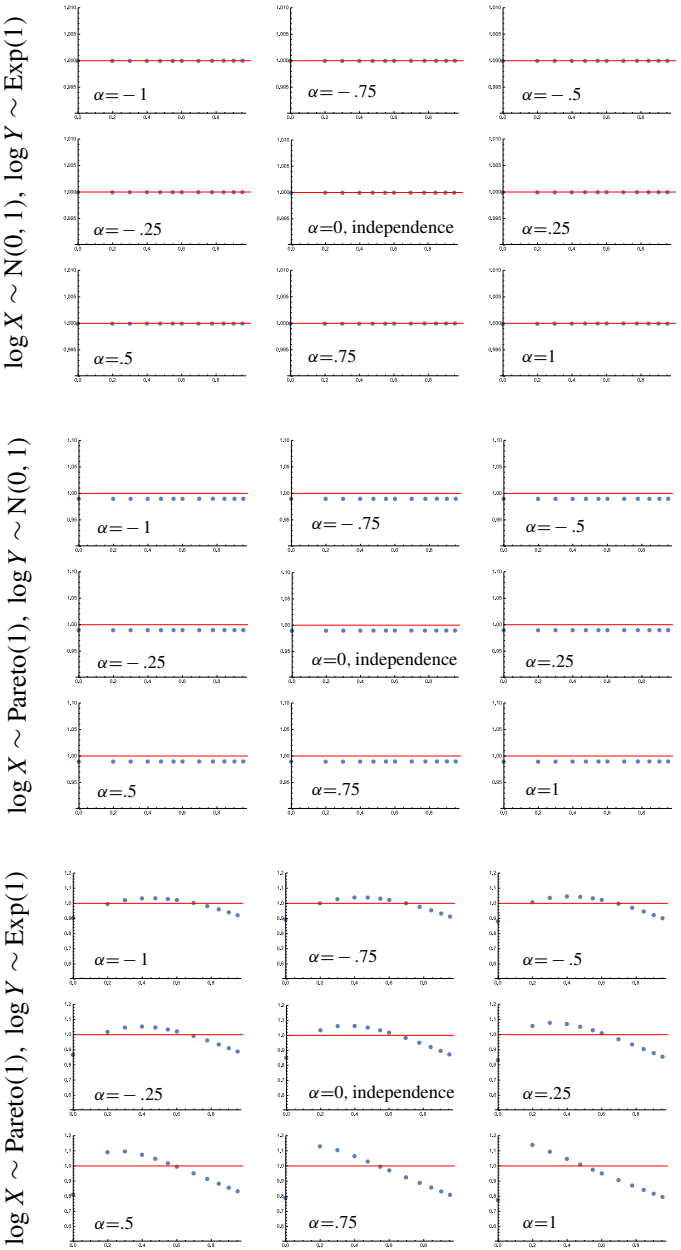

Figure 1. The Ali–Mikhail–Haq 2-copula (see Definition 2.12) modeled on three different sets of marginals with varying dependence parameter $\alpha \in [-1, 1)$. The y-axes of these plots represent the approximate values of the copula PDF of $\log_{10} XY \bmod 1$ at various values of $x \in [0, 1]$, where X and Y are the marginal distributions. The line represents the Benford distribution.

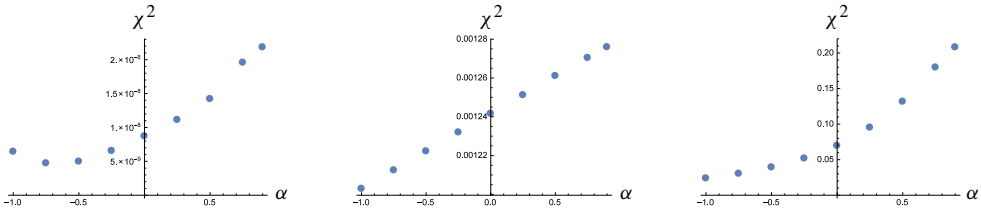

Figure 2. The χ^2-values associated to the plots in Figure 1 for the Ali–Mikhail–Haq copula for pairings (A), (B) and (C), shown from left to right. Each shows the comparison to Benford behavior as α increases. We have 11 degrees of freedom and a significance level of 0.005, so we reject the hypothesis if the value exceeds 2.6. Clearly, only case (C) comes within one order of magnitude of rejecting the hypothesis, so, in loose terms, it is the only case that "comes close" to rejecting the hypothesis. We have imposed a very strict significance level, but it can be clearly seen that a looser significance level of 0.05, perhaps, would likely cause us to reject case (C) entirely.

Ali–Mikhail–Haq copula. Considering the independence case, $\alpha = 0$, in Figure 1 we note that marginal pairings (A) and (B) have an approximately Benford product when independent. Pairing (C), however, does not. From these plots, it is evident that the Ali–Mikhail–Haq copula displays notably consistent Benford behavior, as each plot remains very close to the independence case as α moves over its full range. This is reinforced by the corresponding plots in Figure 2, which display the χ^2 values of each marginal pairing for each value of alpha. We point out that although each plot indicates a general trend away from Benford behavior (the constant function 1), the values for pairing (A) are all smaller than 10^{-7}, making them effectively 0. Similarly, the values for pairing (B) appear to increase linearly, but they are all of order of 10^{-6}. The values for pairing (B) vary from order 10^{-2} to order 10^{-1}, suggesting that the behavior is both significantly less Benford and more variable than the other two pairings.

Gumbel–Barnett copula. These plots suggest that the Gumbel–Barnett copula undergoes even less change over α than the Ali–Mikhail–Haq copula. For pairings (A) and (B), the range for the plots must be restricted to [0.9999, 1.0001] and [0.995, 1.010], respectively, in order to show any change at all. Pairing (C) is not nearly Benford, so its range is expected to vary (recall that the function described by each plot should integrate to 1 in the continuous case). We note, however, that the value at $s = 0$ in pairing (C) appears to vary over a range of 0.1 as α increases. The χ^2 plots in Figure 4 reinforce this interpretation, as in each case the values vary over a significantly small range.

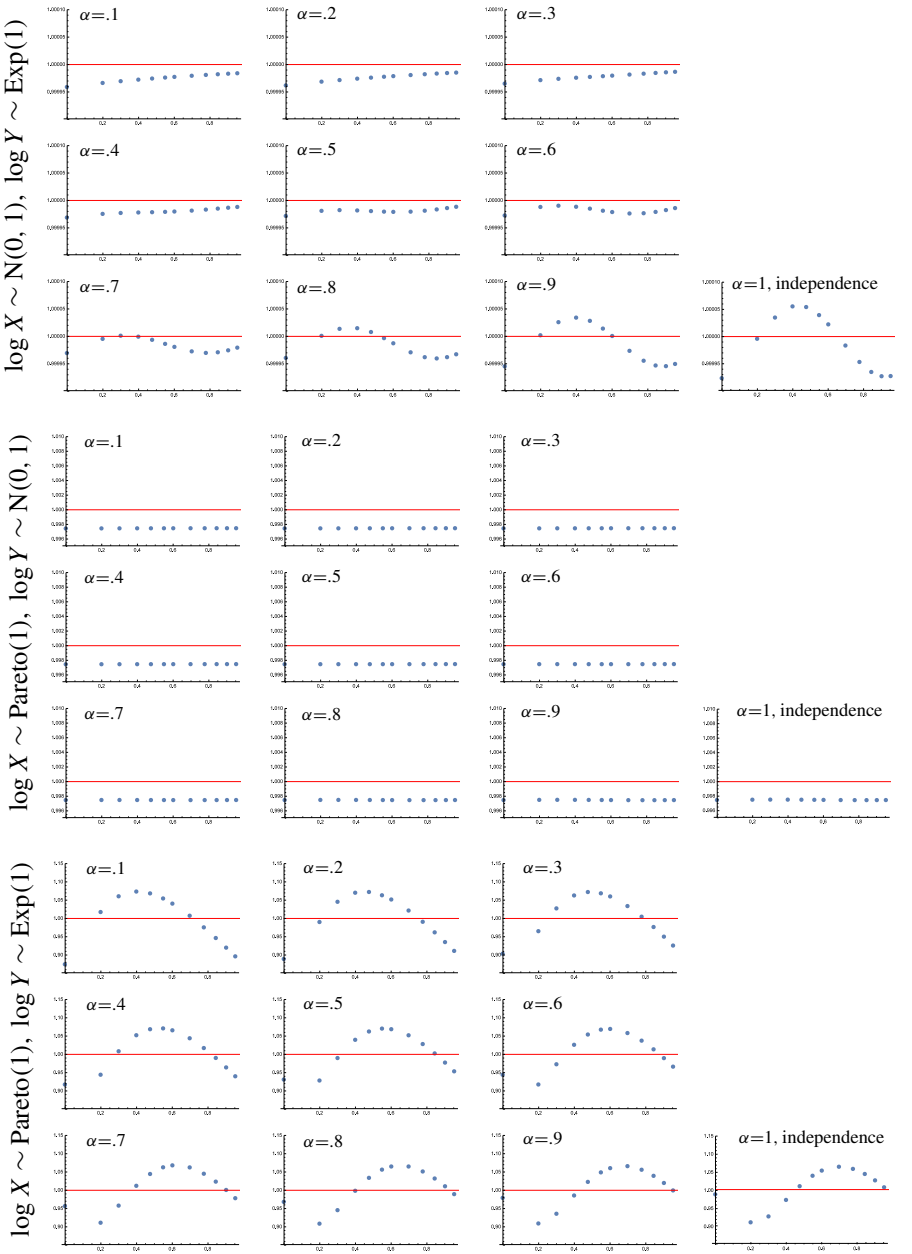

Figure 3. The Gumbel–Barnett 2-copula (see Definition 2.13) modeled on three different sets of marginals with varying dependence parameter $\alpha \in (0, 1]$. The y-axes of these plots represent the approximate values of the copula PDF of $\log_{10} XY \bmod 1$ at various values of $x \in [0, 1]$, where X and Y are the marginal distributions. The line represents the Benford distribution.

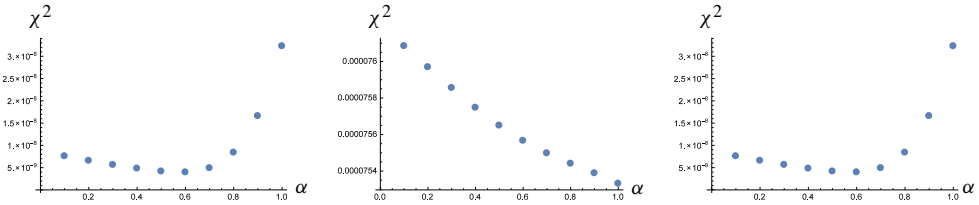

Figure 4. The χ^2-values associated to the plots in Figure 3 for the Gumbel–Barnett copula for pairings (A), (B) and (C), shown from left to right. Each shows the comparison to Benford behavior as α increases. We have 11 degrees of freedom and a significance level of 0.005, so we reject the hypothesis if the value exceeds 2.6. Despite the apparent variation, none of these cases approach the critical value.

This lack of variation is likely due to the actual formula of the copula,

$$C(x, y) = xye^{-\alpha xy}. \tag{4-1}$$

In this case, we have the independence copula, $C(x, y) = xy$ multiplied by a monotonic transformation of the independence copula, e^{-axy}. Thus, it is possible that one or both of these elements serves to preserve the Benford properties of the marginals.

Clayton copula. Unlike the previous two examples, the Clayton copula shows notable variance over α. Although it is not shown here, the independence case for Clayton copulas is $\alpha = 0$. For pairings (A) and (B), it appears that the plots diverge farther and farther away from $y = 1$ as α moves away from 0. For pairing (C), the plots appear to get more random as α grows, and there is no suggestion that Benford behavior may develop as we depart from independence. Furthermore, the plots in Figure 6 show χ^2-values that are significantly higher than those seen for the previous two copulas, suggesting that the dependence imposed by Clayton copula tends to heavily alter any Benford behavior of the marginals.

The results from these three copulas suggest that the preservation of Benford behavior relies more heavily on the underlying structure of the copula than on the Benford behavior of the marginals. Both the Ali–Mikhail–Haq copula and the Gumbel–Barnett copula formulas contain the independence copula, $C(x, y) = xy$. The Clayton copula, however, does not contain the independence copula and is also the only copula of the three to show noticeable variation as the dependence parameter changes.

n-copulas. The previous results suggest that the underlying copula structure has a strong influence on the Benford behavior of 2-copulas. Thus the logical next step is to investigate whether this holds true as we increase the number of marginals.

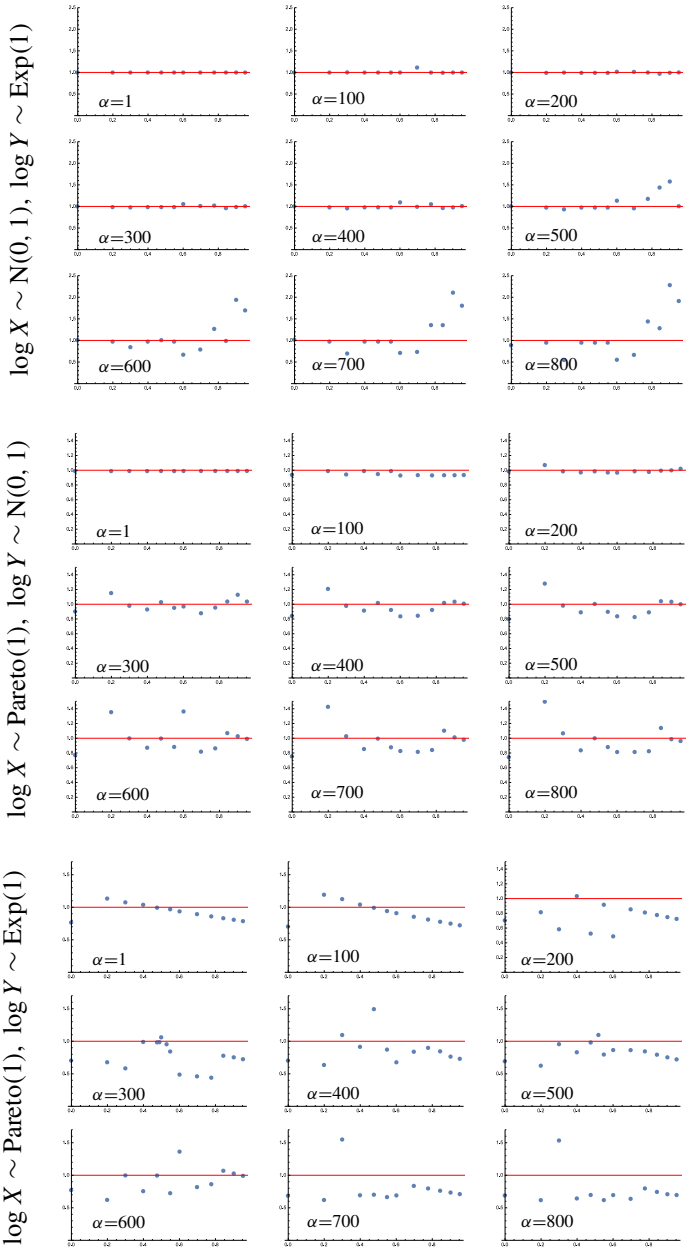

Figure 5. The Clayton 2-copula (see Definition 2.13) modeled on three different sets of marginals with varying dependence parameter $\alpha \in (0, 1]$. The y-axes of these plots represent the approximate values of the copula PDF of $\log_{10} XY \bmod 1$ at various values of $x \in [0, 1]$, where X and Y are the marginal distributions. The line represents the Benford distribution.

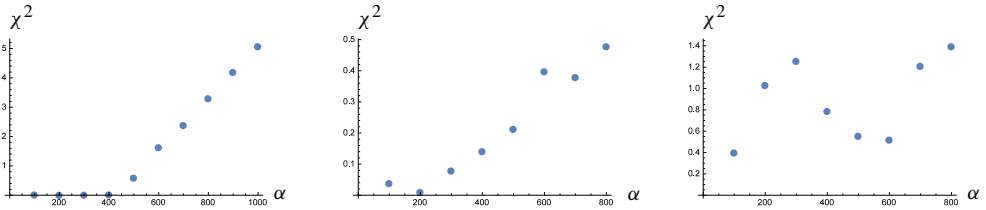

Figure 6. The χ^2 values associated to the plots in Figure 5 for the Clayton copula for pairings (A), (B) and (C), shown from left to right. Each shows the comparison to Benford behavior as α increases. We have 11 degrees of freedom and a significance level of 0.005, so we reject the hypothesis if the value exceeds 2.6. Unlike the previous two copulas, only cases (B) and (C) stay below the critical value. However, the behavior of the plots suggests they will quickly surpass the critical value as α continues to increase.

For all χ^2-tests, we have 8 degrees of freedom and again take a significance level of 0.005. In practice, this means we reject the hypothesis if the value exceeds 1.3.

We consider the most stable of the three previous copulas, the Gumbel–Barnett copula. We fix $\alpha = 0.1$ and set the log, base 10, of all marginals to be identically distributed according to the normal distribution with mean 0 and variance 1, our most Benford-like marginal. We then consider cases where the copula has 2 to 7 marginals. We can see from Figure 7 that the Benford behavior of the Gumbel–Barnett copula begins to fall apart as marginals are added. This is in direct contrast to what would be expected from a central-limit-type property, which should become increasingly more uniform as variables are added. This is further reinforced by the χ^2-values in Figure 8 and suggests that the dependence structure imposed by the copula prevents any leveling-off from happening.

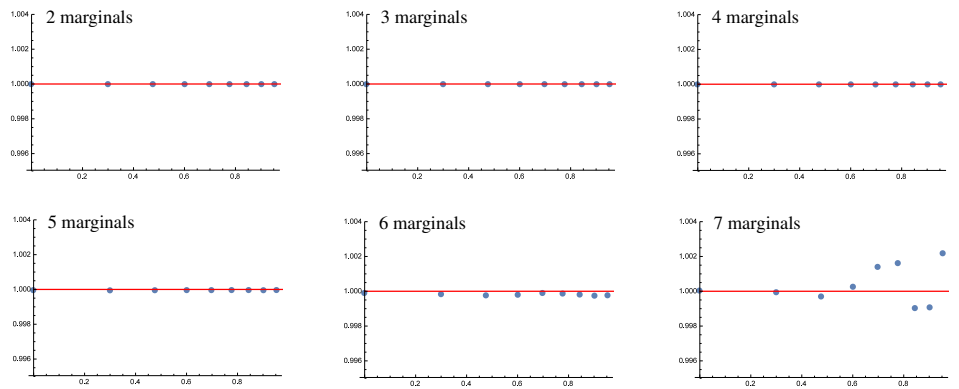

Figure 7. Gumbel–Barnett copula with two to seven marginals.

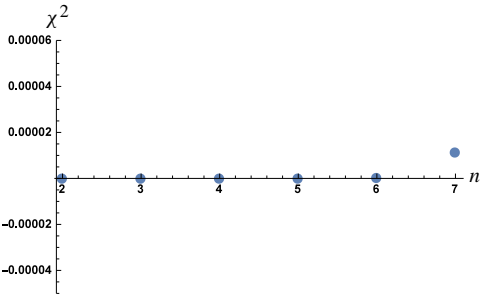

Figure 8. The χ^2-values comparing the behavior of the product to a Benford PDF as the number of marginals increases. We have 8 degrees of freedom and a significance level of 0.005, so we reject the hypothesis if the value exceeds 1.3.

5. Benford distance

Now that we know that we can test for Benford behavior of a product, regardless of dependence, it would be prudent to know how often this behavior is expected to show up. In order to do this, we investigate if the absorptive property of Benford products is common in dependent random variables, or if its presence relies on some sort of proximity to independence.

To get an idea of this, let $\mathcal{W}$ be the space of all n-tuples of continuous random variables $(X_1, X_2, \ldots, X_n)$ for which at least one is Benford. Now let us assume that our set of marginals, $(X_1, X_2, \ldots, X_n)$, form an element in $\mathcal{W}$. Then we know that their product, assuming independence, will always be Benford.

From this, we can restrict our Benford distance, (3-3), to $\mathcal{W}$ and define it as

$$\epsilon_{s,W} = \left| \sum_{k=-\infty}^{\infty} \int_{u_1=-\infty}^{\infty} \cdots \int_{u_{n-1}=-\infty}^{\infty} (N(u_1, \ldots, u_n)|_{\boldsymbol{u}_0}\, du_1 \cdots du_{n-1}) \right|, \tag{5-1}$$

where

$$N(u_1, \ldots, u_n) = 1 - \frac{\partial^n C(F_1(u_1), \ldots, F_{n-1}(u_{n-1}), F_n(u_n))}{\partial u_1 \cdots \partial u_n}$$

and $\boldsymbol{u}_0$ is defined as in Lemma 3.1. Therefore, our problem becomes minimizing the value of $\epsilon_{s,W} = 0$, as proximity to 0 should indicate proximity to a Benford distribution.

Cases that are ϵ away from Benford. Rather than directly calculating the value of $\epsilon_{s,W}$, it may often be more convenient to provide a bound that depends only on the copula C. If the value of $\partial^n C(F_1(u_1), \ldots, F_{n-1}(u_{n-1}), F_n(u_n))/(\partial u_1 \cdots \partial u_n)$ is identically 1 for all values of $(u_1, \ldots, u_n)$, then the value of $\epsilon_{s,W}$ will be identically 0 and our product will be Benford. Even though this case does not

cover all situations in which our product will be Benford, it suggests that a product's distance from Benford may be related to the distance between the function $\partial^n C(F_1(u_1), \ldots, F_{n-1}(u_{n-1}), F_n(u_n))/(\partial u_1 \cdots \partial u_n)$ and the constant function 1. This brings us to the main result of this section.

Theorem 5.1. *Suppose that $X_1, \ldots, X_n$ are continuous random variables where $(X_1, \ldots, X_n) \in \mathcal{W}$. Assume also that they are jointly described by a copula C, where the function*

$$N(u_1, u_2, \ldots, u_n) = 1 - \frac{\partial^n C(F_1(u_1), \ldots, F_{n-1}(u_{n-1}), F_n(u_n))}{\partial u_1 \cdots \partial u_n}$$

is in $L^1(\mathbb{R}^n)$. Let $U_i = \log_B X_i$ for each i and some base, B, and let F_i be the CDFs of U_i for each i. Then the ***L^1 distance from Benford****, defined by*

$$\int_0^1 \left| \sum_{k=-\infty}^{\infty} \int_{u_1=-\infty}^{\infty} \cdots \int_{u_{n-1}=-\infty}^{\infty} (N(u_1, \ldots, u_n)|_{\boldsymbol{u}_0})\, du_1 \cdots du_{n-1} \right| ds \tag{5-2}$$

is bounded above by the L^1 norm of N. In other words

$$\int_0^1 \left| \sum_{k=-\infty}^{\infty} \int_{u_1=-\infty}^{\infty} \cdots \int_{u_{n-1}=-\infty}^{\infty} (1 - N(u_1, \ldots, u_n)|_{\boldsymbol{u}_0})\, du_1 \cdots du_{n-1} \right| ds \leq \|N(u_1, \ldots, u_n)\|_{L^1}. \tag{5-3}$$

We prove this for the two-dimensional case, as the results in n dimensions proceed similarly. We need the following result (see Appendix 1 for a proof).

Lemma 5.2. *Given C_{uv}, $F(u)$, and $G(v)$ as defined before, we have*

$$\|1 - C_{uv}(u, v)\|_{L^1} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(u)g(v)|1 - C_{uv}(F(u), G(v))|\, du\, dv. \tag{5-4}$$

Proof of Theorem 5.1. From the positivity of f and g we have

$$\int_0^1 \left| \sum_{k=-\infty}^{\infty} \int_{-\infty}^{\infty} f(u)g(s+k-u)(1 - C_{uv}(F(u), G(s+k-u)))\, du \right| ds$$

$$\leq \int_0^1 \sum_{k=-\infty}^{\infty} \int_{-\infty}^{\infty} f(u)g(s+k-u)|1 - C_{uv}(F(u), G(s+k-u))|\, du\, ds. \tag{5-5}$$

We investigate exactly what region (5-5) covers. The lines shown in Figure 9 are the sets $A_k = \{(u, v) : v = s + k - u\}$. We integrate

$$f(u)g(s+k-u)(1 - C_{uv}(F(u), G(s+k-u)))$$

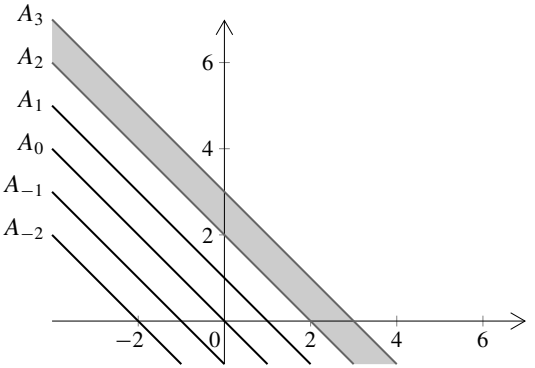

Figure 9. The plane broken up into a few of the sections A_k.

along each of these lines and sum the results over k. The shaded region shows the area covered when A_2 is integrated over s from 0 to 1.

As all of our sums and integrals converge absolutely, by Fubini's theorem we may switch our sum and integral in (5-5) and get

$$\int_0^1 \sum_{k=-\infty}^{\infty} \int_{-\infty}^{\infty} f(u)g(s+k-u)|1-C_{uv}(F(u),G(s+k-u))|\,du\,ds$$

$$= \sum_{k=-\infty}^{\infty} \int_0^1 \int_{-\infty}^{\infty} f(u)g(s+k-u)|1-C_{uv}(F(u),G(s+k-u))|\,du\,ds. \quad (5\text{-}6)$$

From this, we can quickly see that for any k,

$$\int_0^1 \int_{-\infty}^{\infty} f(u)g(s+k-u)|1-C_{uv}(F(u),G(s+k-u))|\,du\,ds \quad (5\text{-}7)$$

is the integral of $f(u)g(s+k-u)|1-C_{uv}(F(u),G(s+k-u))|$ over a region inbetween and including A_k and A_{k+1}, just like the shaded region in Figure 9. Therefore, (5-6) is the sum of the integrals of $f(u)g(s+k-u)|1-C_{uv}(F(u),G(s+k-u))|$ over all of these (disjoint) regions (over all k), which is equivalent to integrating over all of $\mathbb{R}^2$, giving us

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(u)g(v)|1-C_{uv}(F(u),G(v))|\,du\,dv. \quad (5\text{-}8)$$

Finally, from Lemma 5.2, we know that this is equal to $\|1-C_{uv}(u,v)\|_{L^1}$. □

Consequences of an L^1 bound in $\mathbb{R}^2$. What Theorem 5.1 provides is a way to understand the behavior of our probabilities. To see this, let $\mathcal{S} \subset [0,1]$ be the region

over which $\epsilon_{s,W} > \epsilon_N$. If $\epsilon_{s,W}$ is large on $\mathcal{S}$, then the measure of $\mathcal{S}$ must be small in order to conform to (5-3), which requires that if $\|1 - C_{uv}(u, v)\|_{L^1} \leq \epsilon_N$, then $\int_0^1 \epsilon_{s,W}\, ds \leq \epsilon_N$ as well. In fact, the following corollary proves that Theorem 5.1 provides useful information regarding how large $|\mathcal{S}|$ can be.

Corollary 5.3. *Let $\mathcal{S} \subset [0, 1]$ be the set $\{s : \epsilon_{s,W} \geq \epsilon\}$. Then*

$$|\mathcal{S}| \leq \frac{\|1 - C_{uv}(u, v)\|_{L^1}}{\epsilon}. \tag{5-9}$$

Proof. This result comes directly from Markov's inequality:

$$|\{s : \epsilon_{s,W} \geq \epsilon\}| \leq \frac{1}{\epsilon} \int_0^1 \epsilon_{s,W} \leq \frac{\|1 - C_{uv}(u, v)\|_{L^1}}{\epsilon}. \qquad \square$$

6. Applications, future work, and conclusion

Fitting copulas. The results of Section 3 allow us to determine the Benford behavior of the product of n distributions jointly modeled by a specific copula. However, we may wish to go in the other direction and, instead, find a copula that best fits n correlated data sets. Statisticians have several methods for testing the goodness-of-fit to find the best choice of copula in these situations (see [Genest et al. 2006] for some examples and an analysis of several forms of goodness-of-fit tests), but it is not known whether or not these goodness-of-fit tests take Benford behavior into account. That is to say, will the prescribed copula mimic the Benford behavior observed in the data?

The results of Section 4 have shown us that the product of the same set of marginals will not display the same Benford behavior when modeled by different copulas. Thus, Benford behavior is not guaranteed. A natural next step is to investigate how the goodness-of-fit of a copula may or may not be correlated with how well it preserves the expected Benford behavior of the product of two or more marginals. A comparison between the L^1 norm and well-known goodness of fit tests would enable us to see whether or not a strong Benford fit corresponds to a well-fit distribution as a whole. Furthermore, if a stronger Benford fit may be shown to correspond to a smaller L^1 bound, then we may be able to define this bound as a new goodness of fit test for distributions with one or more Benford marginals.

With these results, it is now reasonable to begin searching for specific situations where this analysis of dependence structure would prove useful. As Benford analysis for single-variate distributions has already proven useful in a variety of situations, it is reasonable to assume that the multivariate analysis will be similarly useful. Thus future work may also be directed towards investigating the various applications of these results and how they may be used to improve current practices.

Conclusion. In fields such as actuarial sciences and statistics Benford's law is useful for fraud detection. Furthermore, copulas are a highly effective tool for modeling systems with dependencies. In Section 3 we demonstrated that Benford behavior for dependent variables modeled by a copula may be detected and therefore analyzed to investigate the product of the variables. Thus these results indicate that the Benford's law methods used by professionals on single-variate and/or independent data sets are now at the disposal of individuals who wish to model dependent data via a copula. We then applied these results in Section 4, where we observed that the preservation of Benford behavior appears to rely more heavily on the structure of the copula than on the marginals.

Essentially, the results of Section 3 permit analyses like those carried out in [Cuff et al. 2015; Durst et al. 2016] in which a known distribution, in these cases the Weibull and the inverse-gamma distributions, is analyzed to determine the conditions under which Benford behavior should arise. Once these conditions are established, any non-Benford data set which is expected to come from such a distribution may be considered suspicious enough to warrant a fraud investigation. In the case of copulas, the results of Section 3 allow one to conduct this exact method of analysis on the product of n random variables jointly modeled by a copula C.

Finally, in Section 5 we encountered a useful consequence of considering a distribution's L^1 distance from a Benford distribution to determine a useful bound for this Benford distance. We determined that the Benford distance of a product of n random variables will always be bounded above by the distance between the copula PDF and the class of copulas whose PDFs are identically 1.

Appendix A: Proofs for supporting lemmas and theorems

Lemma 3.1. *Given X and Y positive, continuous random variables with joint distribution modeled by the absolutely continuous copula C, let $U = \log_B X$ and $V = \log_B Y$ for some base, B, and let the (marginal) CDFs of U and V be $F(u)$ and $G(v)$, respectively. Also, let $f(u)$ and $g(v)$ be the PDFs of U and V, respectively. Then*

$$\text{Prob}((U+V) \bmod 1 \le s) = \int_0^s \left(\sum_{k=-\infty}^{\infty} \int_{-\infty}^{\infty} C_{uv}(F(u), G(s+k-u)) f(u) g(s+k-u)\, du \right). \quad \text{(A-1)}$$

Therefore, the PDF of $(U+V) \bmod 1$ is given by

$$\sum_{k=-\infty}^{\infty} \int_{-\infty}^{\infty} C_{uv}(F(u), G(s+k-u)) f(u) g(s+k-u)\, du. \quad \text{(A-2)}$$

Proof. By the invariance of copulas under monotonically increasing functions (Theorem 2.9), we know that the joint CDF of U and V is given by the same copula as X and Y. Thus, the joint CDF of U and V is given by

$$C(F(U), G(V)). \tag{A-3}$$

Then, by definition, the joint PDF of U and V is given by the mixed partial derivative.

$$\begin{aligned}\frac{\partial}{\partial v}\frac{\partial}{\partial u}C(F(u), G(v)) &= C_{uv}(F(u), G(v))f(u)g(v) + C_u(F(u), G(v))\frac{\partial}{\partial v}f(u) \\ &= C_{uv}(F(u), G(v))f(u)g(v). \end{aligned}\tag{A-4}$$

Note that we assume that $du/dv = 0$ since all dependence between U and V is modeled by C.

Note, also, that $\text{Prob}(XY \le 10^s) = \text{Prob}((U+V) \le s)$. Thus we have

$$\begin{aligned}&\text{Prob}((U+V) \bmod 1 \le s) \\ &\quad = \sum_{k=-\infty}^{\infty} \int_{u=-\infty}^{\infty} \int_{v=k-u}^{s+k-u} C_{uv}(F(u), G(v))f(u)g(v)\,dv\,du. \end{aligned}\tag{A-5}$$

If XY is Benford, then (A-5) will equal s for all s. It is, however, easier to test the PDF than the CDF. So we differentiate with respect to s. Let $C_1(u, v)$ be the antiderivative of $C_{uv}(F(u), G(v))f(u)g(v)$ with respect to v. Then

$$\begin{aligned}&\frac{\partial}{\partial s}\sum_{k=-\infty}^{\infty} \int_{u=-\infty}^{\infty} \int_{v=k-u}^{s+k-u} C_{uv}(F(u), G(v))f(u)g(v)\,dv\,du \\ &\quad = \frac{\partial}{\partial s}\sum_{k=-\infty}^{\infty}\left(\int_{u=-\infty}^{\infty}(C_1(u, s+k-u) - C_1(u, k-u))\right)du \\ &\quad = \sum_{k=-\infty}^{\infty}\int_{-\infty}^{\infty} C_{uv}(F(u), G(s+k-u))f(u)g(s+k-u)\,du. \quad \square \end{aligned}$$

Lemma 3.2. *Given U and V, continuous random variables modeled by the copula C with marginals F and G, respectively, there exist a_1, a_2, b_1, and b_2 completely dependent on F or G such that $a_1 < a_2$ and $b_1 < b_2$, and*

$$\begin{aligned}&\sum_{k=-\infty}^{\infty}\int_{-\infty}^{\infty} C_{uv}(F(u), G(s+k-u))f(u)g(s+k-u)\,du \\ &\quad = \sum_{k=b_1}^{b_2}\int_{a_1}^{a_2} C_{uv}(F(u), G(s+k-u))f(u)g(s+k-u)\,du + E_{a,b}(s), \end{aligned}\tag{A-6}$$

where $E_{a,b}(s) \to 0$ *as* $a_1, b_1 \to -\infty$ *and* $a_2, b_2 \to \infty$. *Thus, for any* $\epsilon > 0$, *there exists* $|a_1|$, $|a_2|$, $|b_1|$, *and* $|b_2|$ *large enough such that* $|E_{a,b}(s)| \le \epsilon$.

Proof. Since both the sum and the integral are convergent, the proofs for a_1, a_2 and b_1, b_2 are nearly identical, so we only provide the work here for a_1 and a_2. The same steps may be used in the proof for b_1 and b_2. We also know that $C_{uv}(F(u), G(s+k-u))f(u)g(s+k-u)$ must go to 0 as u goes to $\pm\infty$ because of this convergence. Thus we choose to prove the case where f and/or g converge faster than C_{uv}. If C_{uv} were to converge faster, the results derived here would still suffice. We prove that for any $\epsilon > 0$ we can find a_1 and a_2 such that, for all $u \le a_1$ and all $u \ge a_2$, we have

$$|C_{uv}(F(u), G(s+k-u))f(u)g(s+k-u)| \le \epsilon.$$

Let $\epsilon > 0$, set s and k to be constant, and assume C_{uv} is nonzero everywhere. If C_{uv} is zero at any point, then we have a trivial case. Because F and G are CDFs, we know that $f \to 0$ as $u \to \pm\infty$ and $g \to 0$ as $-u \to \pm\infty$; thus, we may choose a_{f1}, a_{f2}, a_{g1}, and a_{g2} such that, for all $u \le a_{f1}$ and all $u \ge a_{f2}$, we have

$$f(u) \le \sqrt{\frac{\epsilon}{C_{uv}(F(u)G(s+k-u))}}. \tag{A-7}$$

The same can be done for g such that, for all $u \ge a_{g1}$ and all $u \le a_{g2}$, we have

$$g(s+k-u) \le \sqrt{\frac{\epsilon}{C_{uv}(F(u)G(s+k-u))}}. \tag{A-8}$$

Thus, we let $a_1 = \min\{a_{f1}, a_{g1}\}$ and $a_2 = \max\{a_{f2}, a_{g2}\}$. Then, for all $u \le a_1$ and all $u \ge a_2$, we have

$$|C_{uv}(F(u), G(s+k-u))f(u)g(s+k-u)| \le \epsilon. \qquad \square$$

Lemma 5.2. *Given* C_{uv}, $F(u)$, *and* $G(v)$ *as defined in Theorem 5.1, we have*

$$\|1 - C_{uv}(u, v)\|_{L^1} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(u)g(v)|1 - C_{uv}(F(u), G(v))|\,du\,dv. \tag{A-9}$$

Proof. We know that u and v are defined on $[0, 1]$. Thus,

$$\|1 - C_{uv}(u, v)\|_{L^1} = \int_0^1\int_0^1 |1 - C_{uv}(u, v)|\,du\,dv. \tag{A-10}$$

However, by a simple change of variables $u \to F(u)$, $v \to G(v)$ (defined as CDFs, just like before, so their derivatives are $f(u)$ and $g(v)$, both of which are greater

than or equal to 0), we get

$$\|1 - C_{uv}(u, v)\|_{L^1} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(u)g(v)|1 - C_{uv}(F(u), G(v))|\,du\,dv. \qquad \square$$

Appendix B: Computationally testing for Benford behavior

In this section, we use Clayton copulas (see Definition 2.11) to determine the Benford behavior of different combinations of marginals. We specifically look at marginals of the form $X = 10^U$ and $Y = 10^V$, where U and V are N(0, 1) or Exp(1). In all analyses, we let $\alpha = 2$ and $B = 10$. We also provide the independence case for each set of marginals to allow for comparison.

All numerical results and coding were done using Wolfram Mathematica, version 10.1 or later.

Case 1: U and $V \sim N(0, 1)$. Given our definition of X and Y, (3-2) we first determine acceptable values for a_1, b_1, a_2, and b_2 by using an error analysis to test whether or not -10 and 10 should be acceptable values for a_1 and a_2.

We generated a list of the errors caused by truncating the integral at these values for various values of s; the first value of each triple in the list is s, the second is the lower error and the third is the upper error:

```
In[262]:= errorsb =
 Table[{N[Log[10, s]], ea[Log[10, s]], eb[Log[10, s]]}, {s, 1, 9}]

Out[262]= {{0., 6.86784*10^-22, 1.28213*10^-22},
           {0.30103, 9.38169*10^-24, 1.28257*10^-22},
           {0.477121, 3.03058*10^-25, 1.28274*10^-22},
           {0.60206, 2.74232*10^-26, 1.28266*10^-22},
           {0.69897, 4.3443*10^-27, 1.28249*10^-22},
           {0.778151, 9.77379*10^-28, 1.28234*10^-22},
           {0.845098, 2.79567*10^-28, 1.28223*10^-22},
           {0.90309, 9.52164*10^-29, 1.28216*10^-22},
           {0.954243, 3.70245*10^-29, 1.28213*10^-22}}
```

To determine the error caused by truncating the integral, we used the approximation method detailed in Section 3. As the list shows, the error is on the order of 10^{-22} or smaller, indicating that our selections for a_1 and a_2 are good bounds. We took the sum from $k = -20$ to $k = 20$ because we know this will be sufficient, as indicated by the convergence in Figure 10 below.

We now plot in Figure 10 the value of our truncated form of our PDF for different values of s. The line $y = 1$ is included to demonstrate how close to 1 our PDF is for all values of s, suggesting that the product of X and Y with joint PDF modeled by a Clayton copula with $\alpha = 2$ should display Benford behavior.

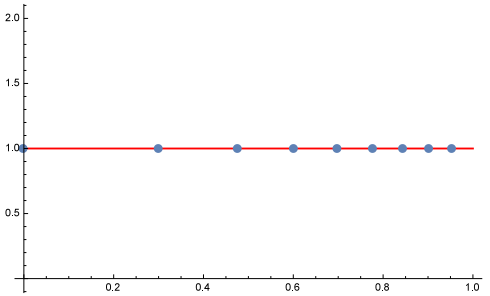

Figure 10. $U \sim \mathrm{N}(0, 1)$ and $V \sim \mathrm{N}(0, 1)$.

Case 2: $U \sim N(0, 1)$ and $V \sim \mathrm{Exp}(1)$. A similar analysis as before was conducted on this new set of variables. Through an identical analysis, we defined the bounds for our integral to be $a = -5$ and $b = 10$, and provide the accumulated errors in the code below, where the first term in each pair is s and the second and third are the lower and upper errors, respectively:

```
In[419]:= Table[{N[Log[10, s]], ea2[Log[10, s]], eb2[Log[10, s]]},
                {s, 1, 9}]

Out[419]= {{0., 3.30411*10^-21, 1.23628*10^-22},
           {0.30103, 2.43577*10^-21, 1.27151*10^-22},
           {0.477121, 2.03887*10^-21, 1.31758*10^-22},
           {0.60206, 1.79746*10^-21, 1.32924*10^-22},
           {0.69897, 1.63021*10^-21, 1.32387*10^-22},
           {0.778151, 1.50526*10^-21, 1.31045*10^-22},
           {0.845098, 1.40717*10^-21, 1.2933*10^-22},
           {0.90309, 1.32741*10^-21, 1.27456*10^-22},
           {0.954243, 1.26084*10^-21, 1.25536*10^-22}}
```

As we can see, the errors are still very, very small.

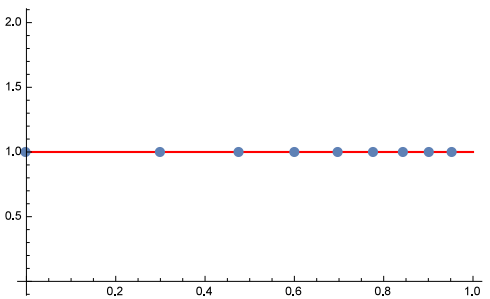

Figure 11. $U \sim N(0, 1)$ and $V \sim \mathrm{Exp}(1)$.

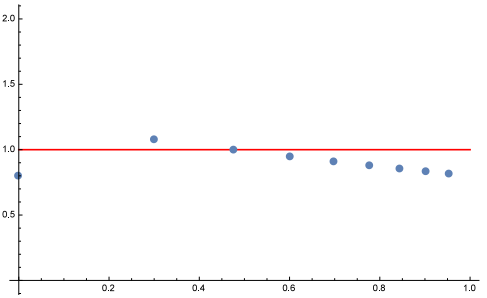

Figure 12. U and $V \sim \text{Exp}(1)$.

We now plot in Figure 11 the value of our truncated form of our PDF for various s. We again note how close the PDF remains to 1 for all values of s, suggesting that the product of X and Y, with joint PDF modeled by a Clayton copula with $\alpha = 2$ should display Benford behavior.

Case 3: $U \sim \text{Exp}(1)$ and $V \sim \text{Exp}(1)$.

Finally, we conduct our analysis on the case of two exponentials. Our error terms for $a = 25$ are generated in the code below (By inspection, we can tell that $C_{uv}(F(u), G(s+k-u))f(u)g(s+k-u)$ will be zero for negative values of u). Again we choose k from 0 to 50, and the first term in each pair is s:

```
In[363]:= Table[{N[Log[10, s]], N[eb1[Log[10, s]]]}, {s, 1, 9}]

Out[363]= {{0., 5.57839*10^-11}, {0.30103, 5.73736*10^-11},
           {0.477121, 5.94524*10^-11}, {0.60206, 5.99786*10^-11},
           {0.698970, 5.97362*10^-11}, {0.778151, 5.91306*10^-11},
           {0.845098, 5.83566*10^-11}, {0.90309, 5.75112*10^-11},
           {0.954243, 5.66447*10^-11}}
```

Now that we know $a = 25$ provides a small enough error, we plot, once again, the PDF for various values of s, as shown in Figure 12. We quickly see that the PDF does not converge to 1 and actually changes for each value of s. Even though we only take our sum out to $k = \pm 50$, this is enough to suggest that Benford behavior is unlikely.

Checking the marginals: To understand why this might be the case, we took a look at the marginal distributions. We note that $X = 10^U$, where $U \sim N[0, 1]$ is a closely Benford distribution with $\chi^2 \approx 0.9918$, but $Y = 10^V$, where $V \sim \text{Exp}[1]$ is not, with $\chi^2 \approx 0.7084$. Thus, in the independent case we would expect that two variables modeled like X, or any product with X, should yield a Benford distribution. The product of two variables modeled like Y, however, should not be Benford.

Acknowledgements

The authors were partially supported by NSF grants DMS-1347804, DMS-1265673 and DMS-1561945, and by Williams College. We thank our colleagues from the Williams College SMALL REU program for many helpful discussions, and the referee for many suggestions which improved the exposition of the paper.

References

[Battersby 2009] S. Battersby, "Statistics hint at fraud in Iranian election", *New Scient.* **202**:2714 (2009), 10.

[Becker et al. 2018] T. Becker, D. Burt, T. C. Corcoran, A. Greaves-Tunnell, J. R. Iafrate, J. Jing, S. J. Miller, J. D. Porfilio, R. Ronan, J. Samranvedhya, F. W. Strauch, and B. Talbut, "Benford's law and continuous dependent random variables", *Ann. Physics* **388** (2018), 350–381. MR Zbl

[Benford 1938] F. Benford, "The law of anomalous numbers", *Proc. Amer. Philos. Soc.* **78**:4 (1938), 551–572. Zbl

[Benford 2009] A. Berger, T. P. Hill, and E. Rogers, "Benford Online Bibliography", database, 2009, available at http://www.benfordonline.net.

[Berger and Hill 2015] A. Berger and T. P. Hill, *An introduction to Benford's law*, Princeton Univ. Press, 2015. MR Zbl

[Cuff et al. 2015] V. Cuff, A. Lewis, and S. J. Miller, "The Weibull distribution and Benford's law", *Involve* **8**:5 (2015), 859–874. MR Zbl

[Durst et al. 2016] R. F. Durst, C. Huynh, A. Lott, S. J. Miller, E. A. Palsson, W. Touw, and G. Vriend, "The inverse gamma distribution and Benford's law", preprint, 2016. arXiv

[Fisher 1997] N. I. Fisher, "Copulas", pp. 159–163 in *Encylopedia of statistical sciences, Update Vol. 1*, edited by S. Kotz et al., Wiley, New York, 1997.

[Fontes and Magni 2004] F. A. C. C. Fontes and L. Magni, "A generalization of Barbalat's lemma with applications to robust model predictive control", pp. art. id. 395 in *Proceedings sixteenth International Symposium on Mathematical Theory of Networks and Systems* (Leuven, Belgium, 2004), edited by B. De Moor et al., Katholieke Univ. Leuven, 2004.

[Genest et al. 2006] C. Genest, J.-F. Quessy, and B. Rémillard, "Goodness-of-fit procedures for copula models based on the probability integral transformation", *Scand. J. Statist.* **33**:2 (2006), 337–366. MR Zbl

[Iafrate et al. 2015] J. R. Iafrate, S. J. Miller, and F. W. Strauch, "Equipartitions and a distribution for numbers: a statistical model for Benford's law", *Phys. Rev. E* (3) **91**:6 (2015), art. id. 062138. MR

[Kpanzou 2007] T. A. Kpanzou, "Copulas in statistics", preprint, African Inst. Math. Sci., 2007, available at https://tinyurl.com/kpancop.

[Miller 2015] S. J. Miller (editor), *Benford's law: theory and applications*, Princeton Univ. Press, 2015. MR Zbl

[Nelsen 2006] R. B. Nelsen, *An introduction to copulas*, 2nd ed., Springer, 2006. MR Zbl

[Nigrini 1999] M. J. Nigrini, "I've got your number", *J. Accountancy* **187**:5 (1999), 79–83.

[Nigrini and Mittermaier 1997] M. Nigrini and L. J. Mittermaier, "The use of Benford's law as an aid in analytical procedures", *Auditing* **16**:2 (1997), 52–67.

[Raimi 1976] R. A. Raimi, "The first digit problem", *Amer. Math. Monthly* **83**:7 (1976), 521–538. MR Zbl

[Singleton 2011] T. W. Singleton, "Understanding and applying Benford's law", *ISACA J.* **3** (2011), 1–4.

[Tao 2010] T. Tao, *An epsilon of room, II: Pages from year three of a mathematical blog*, Grad. Studies in Math. **117**, Amer. Math. Soc., Providence, RI, 2010. MR Zbl

[Wu et al. 2007] F. Wu, E. Valdez, and M. Sherris, "Simulating from exchangeable Archimedean copulas", *Comm. Statist. Simulation Comput.* **36**:5 (2007), 1019–1034. MR Zbl

Received: 2019-01-16 Revised: 2019-04-11 Accepted: 2019-04-15

rfd1@williams.edu *Department of Mathematics and Statistics, Williams College, Williamstown, MA, United States*

Current address: *Division of Applied Mathematics, Brown University, Providence, RI, United States*

sjm1@williams.edu *Department of Mathematics and Statistics, Williams College, Williamstown, MA, United States*

Current address: *Department of Mathematics, Carnegie Mellon University, Pittsburgh, PA, United States*

INVOLVE 12:7 (2019)
dx.doi.org/10.2140/involve.2019.12.1219

Closed geodesics on doubled polygons

Ian M. Adelstein and Adam Y. W. Fong

(Communicated by Frank Morgan)

We study $1/k$-geodesics, those closed geodesics that minimize on any subinterval of length L/k, where L is the length of the geodesic. We investigate the existence and behavior of these curves on doubled polygons and show that every doubled regular n-gon admits a $1/(2n)$-geodesic. For the doubled regular p-gons, with p an odd prime, we conjecture that $k = 2p$ is the minimum value for k such that the space admits a $1/k$-geodesic.

1. Introduction

Traders and explorers have long sought shorter paths across our globe. Columbus in the fifteenth century thought it was possible to reach the East by sailing west. Alas, a continent stood in the way, and in the nineteenth century many explorers searched for the elusive Northwest Passage, a sea route connecting the Atlantic and Pacific via the Arctic Ocean. With the advent of air travel more direct routes became possible; planes often follow the shortest path between two points on the globe. In flat Euclidean space (like the xy-plane) the shortest path between any two points is a straight line. On a sphere the shortest paths are great circles, those curves of intersection between the surface of the sphere and a plane containing its center. This is why when you fly between cities in the northern hemisphere your route travels north towards the pole (see Figure 1).

A geodesic is a locally length-minimizing curve; it is the shortest path between any pair of sufficiently close points on the curve. In flat Euclidean space the geodesics are straight lines. We note that these geodesics are not only locally length-minimizing, but also globally length-minimizing; the straight line is the shortest path between any pair of points on the line, regardless of how close they are. In this paper we study geodesics that fail to minimize globally. As a first example of such a curve consider the geodesic in Figure 2. Another important class of geodesics that fail to minimize globally are the closed geodesics, those geodesics that close up on themselves after finite time.

MSC2010: 53C20, 53C22.
Keywords: closed geodesics, regular polygons, billiard paths.

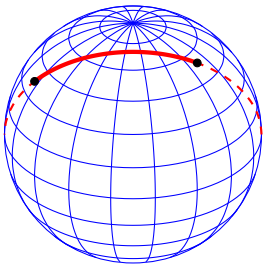

Figure 1. Great circle on a sphere showing the shortest path.

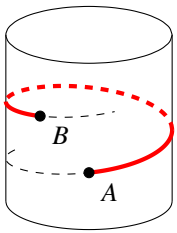

Figure 2. Geodesic on the cylinder that is not the shortest path between points A and B.

Definition 1.1. We use the symbol S^1 to denote the circle. A closed geodesic is a map $\gamma : S^1 \to M$ that is locally length-minimizing at every $t \in S^1$.

The great circles on the sphere are examples of closed geodesics. Fixing any point on the curve, the great circle is the shortest path to every other point on the circle up to its antipodal point, halfway along the length of the curve. If we traverse past the antipodal point, then a shorter path can be found by traversing the circle in the opposite direction, demonstrating that the great circles are not globally length-minimizing. Indeed, every closed geodesic fails to be globally length-minimizing, as traversing in the opposite direction always guarantees a shorter path to points beyond the halfway point.

It is not the case that a closed geodesic will always be the shortest path between pairs of points halfway along the curve. In Figure 3 we see an example of a closed geodesic on a flat torus (the red curve) which does not minimize between pairs of points that are half the length apart. Indeed, the green (dashed) curve provides a shorter path between p and s. Logically, this poses the question of the largest interval on which a given closed geodesic minimizes. To examine this, Sormani [2007, Definition 3.1] introduced the notion of a $1/k$-geodesic.

Definition 1.2. A $1/k$-geodesic is a constant-speed closed geodesic $\gamma : S^1 \to M$ which minimizes on all subintervals of length L/k, where L is the length of the geodesic and $k \in \mathbb{N}$.

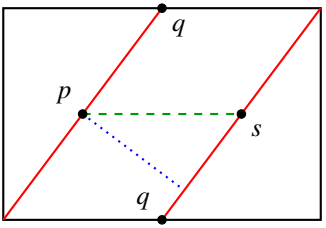

Figure 3. Closed geodesic on a flat torus.

Note that the great circles on the sphere are $\frac{1}{2}$-geodesics, or half-geodesics. The curve in Figure 3 is a $\frac{1}{4}$-geodesic, as it minimizes between all points at length $L/4$ (for example, between the points p and q). The curve does not minimize beyond points at length $L/4$, as is evidenced by the blue (dotted) curve between p and a point on the geodesic beyond q. See also [Adelstein 2016a; 2016b; Ho 2008; Sormani 2007] for more on $1/k$-geodesics. An important first fact about $1/k$-geodesics is that they are as ubiquitous as closed geodesics.

Proposition 1.3 [Sormani 2007, Theorem 3.1]. *Every closed geodesic is a $1/k$-geodesic for some $k \geq 2$.*

Proof. Let $\gamma : S^1 \to M$ be a constant-speed closed geodesic. Then by the local length-minimization property of γ we have for every $t \in S^1 = [0, 2\pi]$ that there exists an $\epsilon_t > 0$ such that γ minimizes on the interval $(t-\epsilon_t, t+\epsilon_t)$. These intervals form an open cover of S^1 and by compactness of the circle we can choose a finite subcover. Let ϵ be the Lebesgue number of the finite subcover, and by the Archimedean property choose $k \geq 2\pi/\epsilon$. Then γ minimizes on all parameter intervals $(t - \pi/k, t + \pi/k)$ and hence γ minimizes on all subintervals of length L/k. □

2. The over-under curve on doubled polygons

We proceed by studying $1/k$-geodesics on doubled regular n-gons. We define a doubled regular n-gon, denoted by X_n, to be the metric space obtained by gluing two regular n-gons along their common edges. We think of the doubled regular n-gons as having a top face and bottom face, so that traversal from one face to the other is possible only by crossing through a point along the shared edges or vertices of the faces. The distance between any two points lying on the same face is the standard Euclidean distance, whereas the distance between two points $x, y \in X_n$ lying on opposite faces is given by $\min_z\{d(x, z)+d(z, y)\}$, where d is the Euclidean distance function on each face and the minimum is taken over all edge points $z \in X_n$.

We next need to determine the behavior of geodesics on these doubled polygons. On any given face the space is Euclidean and the geodesics are straight lines; if two points are on the same face the straight line path between them is a geodesic.

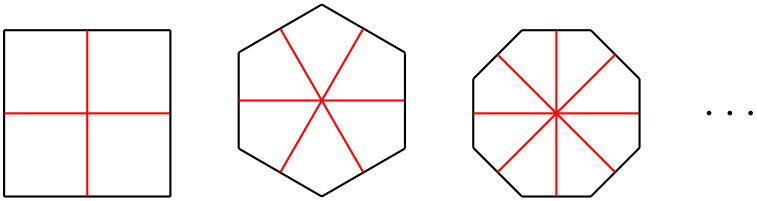

Figure 4. The $n/2$ half-geodesics on X_n, n even. Note that we only depict one face of the doubled polygon, and that these geodesics are the concatenation of straight line paths on the top and bottom faces.

If two points are on opposite faces, a geodesic connecting them must consist of a straight line segment on each face, connected via a shared edge or vertex point. If this geodesic traverses an edge, we can reflect the doubled polygon over this edge, creating a Euclidean space, and conclude that the geodesic on this reflected space must be a straight line. Upon reflecting back over the edge, we see that the angle of incidence is equal to the angle of reflection, i.e., that the geodesics billiard around the edges of the doubled polygons; see [Veech 1992]. An application of Heron's solution to the shortest path problem illuminates this billiard behavior. We also have the following lemma.

Lemma 2.1 [Adelstein 2016a, Lemma 2.1]. *Geodesics on doubled regular n-gons do not contain vertices as interior points.*

Proof. By contradiction assume that the geodesic contains a vertex point. Because regular polygons are convex, we can always reflect the doubled polygon over one of the edges adjacent to the vertex (as in the paragraph above) such that the geodesic in the resulting Euclidean space is kinked with an acute angle. Choosing a pair of geodesic points on either side of the vertex, and considering the triangle formed in the resultant Euclidean space from these two points and the vertex, we conclude via the triangle inequality that there exists a shorter path connecting these points. This contradicts the local length-minimizing property of the geodesic at the vertex. □

The closed geodesics on the doubled regular polygons are interesting to study because of their simplicity. Our research is motivated by the following result:

Proposition 2.2 [Adelstein 2016a, Proposition 2.5]. *Let X_n be a doubled regular n-gon*:

(1) *If n is odd then X_n has no half-geodesics.*

(2) *If n is even then X_n has exactly $n/2$ half-geodesics*: *those curves which pass through the center of each face and perpendicularly through parallel edges*; *see Figure 4.*

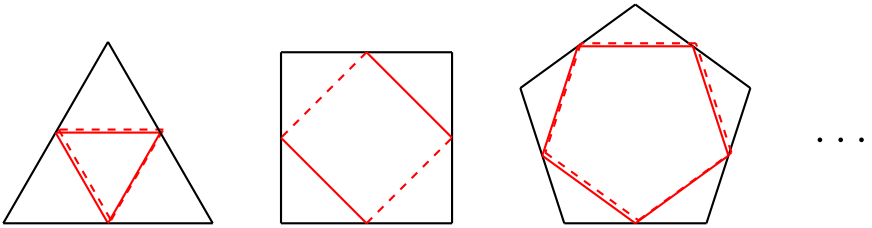

Figure 5. Over-under curves on X_n. Note that we now depict as solid the segments of the geodesic on the top face, and as dashed the segments on the bottom face.

For n odd, the result states that X_n admits no half-geodesics. This naturally leads to the question of the smallest $k \in \mathbb{N}$ such that X_n admits a $1/k$-geodesic. To examine this question we introduce the notion of an over-under curve on X_n.

Definition 2.3. Let $\gamma : S^1 \to X_n$ be the closed geodesic on the doubled regular n-gon that passes through the midpoints of adjacent edges of X_n. We call γ an *over-under* curve between adjacent edges on X_n.

If γ is an over-under curve and $\gamma(t_0)$, $\gamma(t_1)$, and $\gamma(t_2)$ are edge points of X_n with the edge containing $\gamma(t_1)$ adjacent to the edges containing $\gamma(t_0)$ and $\gamma(t_2)$ then the following facts are immediate:

(1) $\gamma|_{(t_0,t_1)}$ and $\gamma|_{(t_1,t_2)}$ are on opposite faces of X_n.

(2) For every $t \in (t_0, t_1)$ and $s \in (t_1, t_2)$ the minimum path between $\gamma(t)$ and $\gamma(s)$ through the edge containing $\gamma(t_1)$ passes through the point $\gamma(t_1)$.

The over-under curves on X_n exhibit distinct behavior depending on the parity of n. If n is even, the curves close smoothly after n segments. If n is odd, the curves close after n segments, but not smoothly. The first and n-th segments are on the same face of X_n, thus forming a corner when they meet at an edge. The curve needs $2n$ segments before closing smoothly, so that the first and $2n$-th segments are on opposite faces (see Figure 5). The following theorem states that the minimizing index of the over-under curves equals the number of segments.

Theorem 2.4. *Let $\gamma : S^1 \to X_n$ be an over-under curve between adjacent edges on a doubled regular n-gon*:

(1) *If n is even then γ is a $1/n$-geodesic.*

(2) *If n is odd then γ is a $1/(2n)$-geodesic.*

Proof. We prove the theorem for n even and note that the proof of the odd case is equivalent after a reparametrization of the curve. Start by parametrizing γ by a circle of length 2π so that each edge point is given by $p_i = \gamma(2\pi i/n)$. To prove the theorem we show that γ is the minimizing path between any pair of points

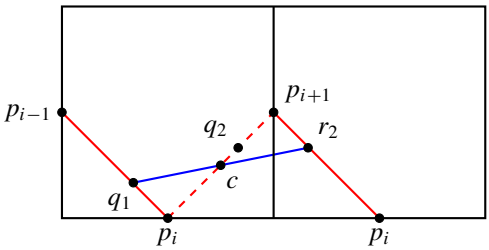

Figure 6. The over-under curve on X_4.

$q_1 = \gamma(t)$ and $q_2 = \gamma(t + 2\pi/n)$. First note that if the q_j are edge points then γ is indeed the minimizing path, as γ is a straight line path on a single face of X_n. Otherwise the q_j are on opposite faces and the segment of γ connecting the pair contains an edge point p_i. Any shorter path between the q_j must cross an edge distinct from the edge containing p_i. It is only necessary to consider paths through the edges containing $p_{i\pm1}$ as we can easily provide a lower bound of $l(\gamma)/n$ for the length of paths through other edges. Without loss of generality we consider only those paths through the edge containing p_{i+1}.

By reflecting the doubled polygon over the edge containing p_{i+1} and considering the top and bottom faces as part of the same plane, we are able to complete the proof in the Euclidean setting. Assume q_1 is on the top face and let r_2 denote the reflection of q_2 through the edge containing p_{i+1} (see Figure 6). We show that the straight-line path between q_1 and r_2 has length at least $l(\gamma)/n$. Let c be the point of intersection between the line segments $\overline{q_1 r_2}$ and $\overline{p_i p_{i+1}}$. Consider the pair of triangles $\triangle q_1 c p_i$ and $\triangle r_2 c p_{i+1}$. By construction we have that the sides opposite $\angle c$ in each triangle have equal length so that applying law of sines to both triangles yields

$$\frac{\sin(\angle q_1)}{Q_1} = \frac{\sin(\angle p_i)}{P_i} = \frac{\sin(\angle c)}{C} = \frac{\sin(\angle r_2)}{R_2} = \frac{\sin(\angle p_{i+1})}{P_{i+1}},$$

where we have used a capital letter to denote the length of the side opposite its angle. We note that $\angle p_i = \pi - \angle p_{i+1}$ so that $\angle r_2 = \angle p_i - \angle c$ and

$$\frac{\sin(\pi - \angle c - \angle p_i)}{Q_1} = \frac{\sin(\angle p_i)}{P_i} = \frac{\sin(\angle p_i - \angle c)}{R_2} = \frac{\sin(\pi - \angle p_i)}{p_{i+1}}.$$

Via the trigonometric identity $\sin(\pi - x) = \sin(x)$ we have $P_i = P_{i+1}$ and

$$\frac{Q_1 + R_2}{2P_i} = \frac{\sin(\angle p_i - \angle c) + \sin(\angle p_i + \angle c)}{2\sin(\angle p_i)} = \frac{2\sin(\angle p_i)\cos(\angle c)}{2\sin(\angle p_i)} = \cos(\angle c) \leq 1.$$

We have therefore shown that $2P_i = P_i + P_{i+1} \geq Q_1 + R_2 = l(\gamma)/n$ and conclude that γ minimizes on all subintervals of length $l(\gamma)/n$. □

3. Bounding the minimizing index

We have shown for n odd that X_n admits a $1/(2n)$-geodesic by explicitly constructing such curves. We now consider whether these curves realize the optimal minimizing property on X_n, i.e., if $k = 2n$ is the smallest $k \in \mathbb{N}$ for which X_n (n odd) admits a $1/k$-geodesic. To quantify this notion Sormani introduced the minimizing index.

Definition 3.1 [Sormani 2007, Definition 3.3]. The minimizing index of a metric space M, denoted by $\operatorname{minind}(M)$, is the smallest $k \in \mathbb{N}$ such that the metric space admits a $1/k$-geodesic.

For n odd the results of the previous section give an upper bound of $2n$ on $\operatorname{minind}(X_n)$. Furthermore, we have seen that such X_n do not admit half-geodesics and consequently that $2 < \operatorname{minind}(X_n) \leq 2n$. A natural question is whether we can sharpen this bound on the minimizing index of X_n. Given a doubled prime-gon it is compelling to believe that its minimizing index is $2p$.

Conjecture 3.2. *If p is an odd prime, then* $\operatorname{minind}(X_p) = 2p$.

Observe here that the primality of p is necessary, since if we have $n = kp$ with $k \geq 2$, we can construct a $1/(2p)$-geodesic by creating an over-under curve between the midpoints of every k-th edge of X_n. Evidence towards this conjecture begins with the following:

Proposition 3.3. *The conjecture is true for the case $p = 3$; i.e., the minimizing index of the doubled regular triangle is 6.*

Proof. We first define the *period* of a closed geodesic on a doubled polygon to be its total number of segments. As these geodesics must close smoothly, we have that the period is always even. Also note because a geodesic on a doubled polygon will never minimize on an open segment that contains multiple edge points, the period provides a lower bound on the minimizing index of a geodesic (the smallest $k \in \mathbb{N}$ such that it is a $1/k$-geodesic).

We have therefore reduced the problem to showing that those closed geodesics with period less than 6 have minimizing index at least 6. We have already established that X_3 does not admit half-geodesics, and that the period must be even, so we need only consider those closed geodesics with period 4. Such curves can be classified: they must leave an edge with angle $\frac{\pi}{6}$, traverse an adjacent edge perpendicularly, return to the starting edge (at the same point, but not with the same velocity), traverse the remaining edge perpendicularly, and return to the starting point to close up smoothly (see Figure 7).

It remains to show that any period-4 geodesic on X_3 has minimizing index at least 6. We first show that the period-4 geodesic from Figure 7 has minimizing index

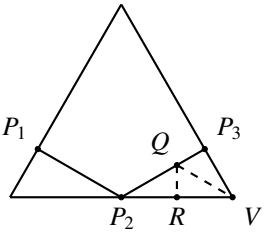

Figure 7. Closed geodesic on X_3 with period 4 and minimizing index 6.

at least 6. In this figure QV is the bisector of angle V and QR is perpendicular to VP_2. Using properties of similar triangles we have

$$|QR| = |QP_3| = \frac{|P_2P_3|}{3} = \frac{L}{12}.$$

This demonstrates that there exist two equal-length paths between Q and its corresponding point on the bottom face: one along our geodesic through P_3, and another through R. The geodesic therefore cannot minimize beyond this segment of length $L/6$, and we conclude that the minimizing index must be at least 6. For a period-4 geodesic on X_3 that does not contain the midpoint of an edge, a similar argument shows that the minimizing index must be strictly greater than 6. □

Please note that Proposition 3.3 did not appear in the original version of this paper. The proof was sketched by the undergraduate research group [Adelstein et al. 2019] and independently by one of the referees (who also produced Figure 7). The original paper had an argument equivalent to the last paragraph of the proof showing that the minimizing index of the geodesic from Figure 7 is at least 6, but did not classify all period-4 geodesics, and therefore did not determine the minimizing index of X_3.

It is reasonable to believe that a similar argument could be used to show that $\text{minind}(X_5) = 10$. It need only be shown that closed geodesics of period 4, 6, or 8 have minimizing index at least 10. One quickly realizes that this direction of reasoning will prove untenable for resolving the conjecture; as p grows it becomes prohibitively difficult to complete such an analysis. As a partial solution to the conjecture we present the following:

Theorem 3.4 [Adelstein et al. 2019, Theorem 2]. *For p prime, as $p \to \infty$, the minimizing index of X_p grows without bound.*

This theorem was proved after the completion of this paper by a subsequent undergraduate research group [Adelstein et al. 2019]. The proof involves a careful study of the closed geodesics on doubled polygons, developing new techniques to study their minimizing properties. To the best of our knowledge Conjecture 3.2 remains open, and we invite the reader to pursue their own investigations.

Acknowledgements

The authors would like to thank the Faculty Research Committee at Trinity College for funding Fong's on-campus research with Adelstein through the Student Research Program. We also acknowledge the wonderful work of Brett C. Smith who recreated all the figures in this paper for publication.

References

[Adelstein 2016a] I. M. Adelstein, "Existence and nonexistence of half-geodesics on S^2", *Proc. Amer. Math. Soc.* **144**:7 (2016), 3085–3091. MR Zbl

[Adelstein 2016b] I. M. Adelstein, "Minimizing closed geodesics via critical points of the uniform energy", *Math. Res. Lett.* **23**:4 (2016), 953–972. MR Zbl

[Adelstein et al. 2019] I. Adelstein, A. Azvolinsky, J. Hinman, and A. Schlesinger, "Minimizing closed geodesics on polygons and disks", preprint, 2019. arXiv

[Ho 2008] W. K. Ho, "Manifolds without $1/k$-geodesics", *Israel J. Math.* **168** (2008), 189–200. MR Zbl

[Sormani 2007] C. Sormani, "Convergence and the length spectrum", *Adv. Math.* **213**:1 (2007), 405–439. MR Zbl

[Veech 1992] W. A. Veech, "The billiard in a regular polygon", *Geom. Funct. Anal.* **2**:3 (1992), 341–379. MR Zbl

Received: 2019-01-24 Revised: 2019-02-07 Accepted: 2019-02-18

iadelstein@gmail.com *Department of Mathematics, Yale University, New Haven, CT, United States*

adam.y.w.fong@gmail.com *Department of Mathematics, Trinity College, Hartford, CT, United States*

INVOLVE 12:7 (2019)
dx.doi.org/10.2140/involve.2019.12.1229

Sign pattern matrices that allow inertia $\mathbb{S}_n$

Adam H. Berliner, Derek DeBlieck and Deepak Shah

(Communicated by Chi-Kwong Li)

Sign pattern matrices of order n that allow inertias in the set $\mathbb{S}_n$ are considered. All sign patterns of order 3 (up to equivalence) that allow $\mathbb{S}_3$ are classified and organized according to their associated directed graphs. Furthermore, a minimal set of such matrices is found. Then, given a pattern of order n that allows $\mathbb{S}_n$, a construction is given that generates families of irreducible sign patterns of order $n+1$ that allow $\mathbb{S}_{n+1}$.

1. Introduction

The *inertia* of a real matrix A of order n is an ordered triple $\mathrm{i}(A) = (n_+, n_-, n_0)$ of nonnegative integers summing to n, where n_+, n_-, n_0 are the number of eigenvalues of A with positive, negative, and zero real parts, respectively.

A *sign pattern matrix* is a matrix $\mathcal{A}$ of order n with entries in $\{+, -, 0\}$. The set $Q(\mathcal{A})$ denotes the set of all real-valued matrices A with corresponding sign pattern $\mathcal{A}$. Alternatively, we say that $A \in Q(\mathcal{A})$ is a *realization* of $\mathcal{A}$. If $\mathcal{A}$ is a sign pattern of order n, then we say that $\mathcal{A}$ has inertia $\mathrm{i}(\mathcal{A}) = \{\mathrm{i}(A) : A \in Q(\mathcal{A})\}$.

In a dynamical system, the presence of a zero eigenvalue of the Jacobian matrix at an equilibrium may signal onset of instability. Varying a parameter may move eigenvalues from all having negative real parts to having a simple zero eigenvalue, which then moves to have a positive real part, while the other eigenvalues maintain negative real parts. Thus, the inertia begins at $(0, n, 0)$, and with parameter variation, changes to $(0, n-1, 1)$ and then finally to $(1, n-1, 0)$.

With this motivation in mind, the inertia set $\mathbb{S}_n$ (for $n \geq 2$) is defined as

$$\mathbb{S}_n = \{(0, n, 0), (0, n-1, 1), (1, n-1, 0)\}.$$

We are particularly interested in studying irreducible sign patterns that *allow* $\mathbb{S}_n$, i.e., $\mathbb{S}_n \subseteq \mathrm{i}(\mathcal{A}_n)$.

Introduced in [Bodine et al. 2012], the *refined inertia* of a matrix $\mathcal{A}$ is the 4-tuple $\mathrm{ri}(A) = (n_+, n_-, n_z, 2n_p)$, where n_+, n_- are defined as before, n_z is the number

MSC2010: primary 15B35, 15A18, 05C50; secondary 05C20.
Keywords: sign pattern, zero-nonzero pattern, inertia, digraph, Jacobian.

of zero eigenvalues, and $2n_p$ is the number of nonzero pure imaginary eigenvalues. Using the notation of refined inertia, $\mathbb{S}_n = \{(0,n,0,0),(0,n-1,1,0),(1,n-1,0,0)\}$. Several results regarding other sets of refined inertias can be found in [Gao et al. 2016a; 2016b; Garnett et al. 2013; 2014]. Many similar techniques and ideas are used in this paper.

For simplicity, we identify sign patterns up to *equivalence*. Any combination of transposition, permutation similarity, and signature similarity leaves the eigenvalues of a matrix unchanged. For our purposes, it is convenient to organize sign patterns by their *associated digraph*. For $\mathcal{A} = [\alpha_{ij}]$ (or a realization $A = [a_{ij}]$) of order n, its associated digraph $D(\mathcal{A})$ is a directed graph on n vertices where there is an arc from vertex i to vertex j if and only if $\alpha_{ij} \neq 0$. Two digraphs are equivalent if and only if their associated zero-nonzero patterns are equivalent via transposition and/or permutation similarity.

In order for a sign pattern $\mathcal{A}$ to be irreducible, the associated digraph $D(\mathcal{A})$ must be strongly connected. A sign pattern $\mathcal{A}$ is *sign singular* if $n_0 > 0$ for all $A \in \mathcal{Q}(\mathcal{A})$ and is *sign-nonsingular* if $n_0 = 0$ for all $A \in \mathcal{Q}(\mathcal{A})$. Thus, in order for $\mathcal{A}$ to allow $\mathbb{S}_n$, $\mathcal{A}$ can neither be sign singular nor sign-nonsingular. In particular, this means that the determinant expansion of $\mathcal{A}$ must have at least two nonzero terms. A nonzero term in the determinant expansion of $\mathcal{A}$ corresponds to the existence of a *generalized n-cycle* in the associated digraph $D(\mathcal{A})$ (that is, a disjoint collection of cycles that use all n vertices of $D(\mathcal{A})$). Furthermore, any sign pattern $\mathcal{A}$ where $\mathrm{i}(A) = (0,n,0)$ must have at least one negative diagonal entry. Thus, for our purposes, we need only consider strongly connected digraphs that contain at least one loop and at least two generalized n-cycles.

For $n = 2$, there are two nonequivalent sign patterns that allow $\mathbb{S}_2$, namely $\left[\begin{smallmatrix} - & - \\ - & - \end{smallmatrix}\right]$ and $\left[\begin{smallmatrix} - & - \\ + & + \end{smallmatrix}\right]$ (see [Olesky et al. 2013]). The first sign pattern requires $\mathbb{S}_2$. The second pattern attains every possible spectrum allowed by a real matrix, and such a pattern is called *spectrally arbitrary*. A sign pattern $\widehat{\mathcal{A}}$ is a *superpattern* of $\mathcal{A}$ if $\mathcal{A}$ can be obtained from $\widehat{\mathcal{A}}$ by changing any number of nonzero entries to 0. In [Berliner et al. 2018], sufficient conditions for a sign pattern and all of its superpatterns to allow $\mathbb{S}_n$ are given. Suppose $A = [a_{ij}]$ is a real matrix of order n having $m \geq n$ nonzero entries and $\mathrm{i}(A) = (0, n-1, 1)$. If the m nonzero entries $a_{i_1,j_1}, \ldots, a_{i_m,j_m}$ are replaced by variables $x_1, \ldots, x_m$ to obtain the matrix X, the characteristic polynomial of X is

$$c_X(z) = z^n + p_1 z^{n-1} + \cdots + p_{n-1} z + p_n,$$

with coefficients $p_1, \ldots, p_n$ depending on $x_1, \ldots, x_m$. The $n \times m$ *Jacobian matrix* J of A has (i,j)-entry equal to $\partial p_i(x_1, \ldots, x_m)/\partial x_j$ evaluated at $(x_1, \ldots, x_m) = (a_{i_1,j_1}, \ldots, a_{i_m,j_m})$. If J has rank n, then A *allows a Jacobian matrix of full rank*. This definition, which uses a rectangular Jacobian matrix as in [Garnett and Shader

2013], is equivalent to the determinantal property that A "allows a nonzero Jacobian" as defined in [Cavers and Vander Meulen 2005]. The following theorem is proved in [Berliner et al. 2018, Theorem 2.2].

Theorem 1.1. *Let $\mathcal{A}$ be an $n \times n$ sign pattern that allows inertia $(0, n-1, 1)$ and let $A \in \mathcal{Q}(\mathcal{A})$ with $\mathrm{i}(A) = (0, n-1, 1)$. If A allows a Jacobian matrix of full rank, then every superpattern $\hat{\mathcal{A}}$ of $\mathcal{A}$ (including $\mathcal{A}$ itself) allows $\mathbb{S}_n$.*

In Section 2, we classify all nonequivalent sign patterns of order 3 that allow $\mathbb{S}_3$. In Section 3, we give a construction that, using a sign pattern of order m that allows $\mathbb{S}_m$, creates sign patterns of any order $n > m$ that allow $\mathbb{S}_n$. This construction allows us to use the sign patterns of order 3 that allow $\mathbb{S}_3$ to create sign patterns of order $n > 3$ that allow $\mathbb{S}_n$.

2. Sign patterns allowing $\mathbb{S}_3$

In this section, we classify all sign patterns of order 3 that allow $\mathbb{S}_3$. First, we may restrict our attention to sign patterns $\mathcal{A}$ whose associated digraph $D(\mathcal{A})$ is strongly connected, has at least one loop, and contains two or more generalized 3-cycles. Without loops included, there are only five nonequivalent strongly connected digraphs of order 3, as shown in Figure 1. For these, we use the same digraph labeling as in [Berliner et al. 2017] (up to equivalence). Adding loops in and enforcing the generalized 3-cycle requirement, we then focus solely on sign patterns associated with the looped digraphs described in Table 1 (again up to equivalence).

If $\mathcal{A}$ is a sign pattern of order 3 having a realization with inertia $(0, 2, 1)$ that allows a Jacobian of full rank, then by Theorem 1.1 any superpattern of $\mathcal{A}$ will

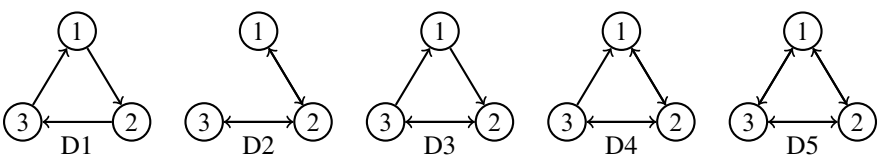

Figure 1. Strongly connected digraphs of order 3.

strongly connected digraph	nonequivalent loop locations
D1	123
D2	13, 123
D3	1, 13, 123
D4	1, 12, 13, 123
D5	1, 13, 123

Table 1. Nonequivalent strongly connected digraphs with two or more generalized 3-cycles.

$$\begin{bmatrix} - & + & 0 \\ 0 & - & + \\ + & 0 & - \end{bmatrix} \begin{bmatrix} + & - & 0 \\ 0 & - & - \\ - & 0 & - \end{bmatrix} \begin{bmatrix} - & - & 0 \\ - & 0 & - \\ 0 & + & - \end{bmatrix} \begin{bmatrix} - & + & 0 \\ - & 0 & + \\ 0 & - & + \end{bmatrix} \begin{bmatrix} - & - & 0 \\ - & - & - \\ 0 & - & - \end{bmatrix} \begin{bmatrix} - & - & 0 \\ - & - & - \\ 0 & + & + \end{bmatrix}$$
$$\begin{bmatrix} - & - & 0 \\ + & + & - \\ 0 & + & - \end{bmatrix} \begin{bmatrix} + & - & 0 \\ + & - & - \\ 0 & - & + \end{bmatrix} \begin{bmatrix} - & + & 0 \\ 0 & 0 & + \\ + & - & 0 \end{bmatrix} \begin{bmatrix} - & + & 0 \\ 0 & 0 & + \\ - & + & - \end{bmatrix} \begin{bmatrix} + & + & 0 \\ 0 & 0 & - \\ + & + & - \end{bmatrix} \begin{bmatrix} - & - & 0 \\ + & 0 & + \\ + & + & 0 \end{bmatrix} \begin{bmatrix} + & + & 0 \\ - & - & + \\ + & + & 0 \end{bmatrix}$$

Figure 2. $\mathbb{S}_3$-minimal sign patterns.

automatically allow $\mathbb{S}_3$. Thus, we will focus on $\mathbb{S}_3$-*minimal sign patterns*, i.e., patterns having a realization with inertia $(0, 2, 1)$ that allows a Jacobian of full rank that are not superpatterns of a pattern having a realization with inertia $(0, 2, 1)$ that allows a Jacobian of full rank.

Of the 200 nonequivalent sign patterns of order 3, 111 allow $\mathbb{S}_3$ and 13 of these are $\mathbb{S}_3$-minimal sign patterns (see Figure 2). Of the $\mathbb{S}_3$-minimal sign patterns, two have associated digraph D1, six are associated to D2, three are associated to D3, and two are associated to D4. All other nonequivalent sign patterns of order 3 that allow $\mathbb{S}_3$ are equivalent to a superpattern of one of these 13, and thus automatically allow $\mathbb{S}_3$. These superpatterns can be found in Appendix A.

Below, we illustrate the method for one of the 13 $\mathbb{S}_3$-minimal sign patterns.

Example 2.1. The sign pattern $\mathcal{A}$ and a realization $A \in Q(\mathcal{A})$ (with associated digraph D2) are given by

$$\mathcal{A} = \begin{bmatrix} - & - & 0 \\ - & 0 & - \\ 0 & + & - \end{bmatrix} \quad \text{and} \quad A = \begin{bmatrix} -1 & -1 & 0 \\ -a & 0 & -1 \\ 0 & b & -c \end{bmatrix},$$

with $a, b, c > 0$. Without loss of generality, one diagonal entry and two other entries in the digraph associated with $\mathcal{A}$ (corresponding to a spanning tree) can be set equal to ± 1. The characteristic polynomial of A is

$$c_A(z) = z^3 + (1 + c)z^2 + (b + c - a)z + (b - ac).$$

In order to realize inertia $(0, 2, 1)$, we must have $b = ac$. If $a = 1$ and $b = c = 2$, then $\mathrm{i}(A) = (0, 2, 1)$, as desired. We now check if A allows a nonzero Jacobian. We replace the nonzero entries of A with variables to get

$$X_A = \begin{bmatrix} x_1 & x_2 & 0 \\ x_3 & 0 & x_4 \\ 0 & x_5 & x_6 \end{bmatrix},$$

which has characteristic polynomial

$$c_{X_A}(z) = z^3 - (x_1 + x_6)z^2 + (x_1x_6 - x_2x_3 - x_4x_5)z + (x_1x_4x_5 + x_2x_3x_6).$$

Calculating the Jacobian matrix of X_A, we get

$$J_{X_A} = \begin{bmatrix} -1 & 0 & 0 & 0 & 0 & -1 \\ x_6 & -x_3 & -x_2 & -x_5 & -x_4 & x_1 \\ x_4x_5 & x_3x_6 & x_2x_6 & x_1x_5 & x_1x_4 & x_2x_3 \end{bmatrix},$$

which when evaluated at $x_1 = x_2 = x_3 = x_4 = -1,\ x_5 = 2,\ x_6 = -2$ has full rank. By Theorem 1.1, $\mathcal{A}$ and all of its superpatterns allow $\mathbb{S}_3$.

The other 89 nonequivalent sign patterns do not allow $\mathbb{S}_3$. Several (57) of these sign patterns do not allow $\mathbb{S}_3$ because they are sign-nonsingular and cannot possibly allow the inertia $(0, n-1, 1)$. These patterns, for $n = 3$, can be found in Appendix B.1. The 32 remaining patterns (see Appendix B.2) are not sign-nonsingular, but nonetheless do not allow inertia $(0, 2, 1)$ for other algebraic reasons. Here, we illustrate the method for one of the sign patterns that is not sign-nonsingular yet does not allow inertia $(0, 2, 1)$.

Example 2.2. The sign pattern $\mathcal{A}$ and a realization $A \in Q(\mathcal{A})$ (with associated digraph D3) are given by

$$\mathcal{A} = \begin{bmatrix} - & - & 0 \\ 0 & 0 & + \\ + & + & + \end{bmatrix} \quad \text{and} \quad A = \begin{bmatrix} -1 & -1 & 0 \\ 0 & 0 & 1 \\ a & b & c \end{bmatrix},$$

with $a, b, c > 0$. Without loss of generality, one diagonal entry and two entries on the 3-cycle in the digraph associated with $\mathcal{A}$ can be set equal to ± 1. The characteristic polynomial of A is

$$c_A(z) = z^3 + (1-c)z^2 + (-b-c)z + (a-b).$$

Since $b, c > 0$, it must be the case that $-b - c < 0$. However, in order to allow inertia $(0, 2, 1)$, the quadratic and linear coefficients of $c_A(z)$ must be able to be simultaneously positive. Thus $\mathcal{A}$ does not allow $\mathbb{S}_3$.

3. The Jacobian and patterns of higher order

We begin with a sign pattern of order n that allows $\mathbb{S}_n$ and give a construction that yields a sign pattern that allows $\mathbb{S}_{n+1}$. If $\mathcal{A}$ is a sign pattern of order n, we consider the $(n+1) \times (n+1)$ sign pattern

$$\mathcal{A}^- = \left[\begin{array}{ccc|c} & & & 0 \\ & \mathcal{A} & & \vdots \\ & & & 0 \\ \hline 0 & \cdots & 0 & - \end{array}\right].$$

Then, $(n_+, n_-, n_0) \in \mathrm{i}(\mathcal{A})$ if and only if $(n_+, n_- + 1, n_0) \in \mathrm{i}(\mathcal{A}^-)$. It follows that $\mathcal{A}^-$ allows $\mathbb{S}_{n+1}$ if and only if $\mathcal{A}$ allows $\mathbb{S}_n$. An analogous result holds for n_+ if

we create the $(n+1)\times(n+1)$ sign pattern $\mathcal{A}^+$ by replacing the lower-right corner entry of $\mathcal{A}^-$ by $+$.

Theorem 3.1. *Let $\mathcal{A}$ and $\mathcal{A}^-$ be defined as above, where $\mathcal{A}$ has at least n nonzero entries. If A is a realization of $\mathcal{A}$ that allows a Jacobian of full rank for which* $\mathrm{i}(A) = (n_+, n_-, n_0)$, *then there exists a realization B of $\mathcal{A}^-$ that allows a Jacobian of full rank and* $\mathrm{i}(B) = (n_+, 1+n_-, n_0)$.

Proof. Let A be a realization of $\mathcal{A}$ that has inertia (n_+, n_-, n_0) and allows a Jacobian of full rank. Then, replacing the $m \geq n$ nonzero entries $a_{i_1,j_1}, \ldots, a_{i_m,j_m}$ of A with variables $x_1, \ldots, x_m$, the characteristic polynomial is

$$p_A(z) = z^n + p_1 z^{n-1} + \cdots + p_{n-1} z + p_n,$$

where $p_1, \ldots, p_n$ are functions of $x_1, \ldots, x_m$. Furthermore, we know the matrix $J_{X_A} = [\partial p_i / \partial x_j]$ has full rank when evaluated at $(x_1, \ldots, x_m) = (a_{i_1,j_1}, \ldots, a_{i_m,j_m})$.

In order to obtain the Jacobian matrix for B, we replace the lower-right corner entry with variable $\hat{x}$ and the other m entries with the same variables $x_1, \ldots, x_m$ as with A. Thus, we obtain the characteristic polynomial

$$\begin{aligned} p_B(z) &= p_A(z)(z-\hat{x}) = (z^n + p_1 z^{n-1} + \cdots + p_{n-1} z + p_n)(z-\hat{x}) \\ &= z^{n+1} + (p_1 - \hat{x}) z^n + (p_2 - \hat{x}\, p_1) z^{n-1} + \cdots + (p_n - \hat{x}\, p_{n-1}) z - \hat{x}\, p_n \end{aligned}$$

and we have

$$J_{X_B} = \begin{bmatrix} \dfrac{\partial p_1}{\partial x_1} & \cdots & \dfrac{\partial p_1}{\partial x_m} & -1 \\ \dfrac{\partial p_2}{\partial x_1} - \hat{x}\dfrac{\partial p_1}{\partial x_1} & \cdots & \dfrac{\partial p_2}{\partial x_m} - \hat{x}\dfrac{\partial p_1}{\partial x_m} & -p_1 \\ \vdots & & \vdots & \vdots \\ \dfrac{\partial p_n}{\partial x_1} - \hat{x}\dfrac{\partial p_{n-1}}{\partial x_1} & \cdots & \dfrac{\partial p_n}{\partial x_m} - \hat{x}\dfrac{\partial p_{n-1}}{\partial x_m} & -p_{n-1} \\ -\hat{x}\dfrac{\partial p_n}{\partial x_1} & \cdots & -\hat{x}\dfrac{\partial p_n}{\partial x_m} & -p_n \end{bmatrix}.$$

We sequentially perform $n-1$ row operations on J_{X_B}, where the i-th row operation adds $\hat{x}$ times row i to row $i+1$. The resulting matrix is

$$J = \begin{bmatrix} \dfrac{\partial p_1}{\partial x_1} & \cdots & \dfrac{\partial p_1}{\partial x_m} & -1 \\ \dfrac{\partial p_2}{\partial x_1} & \cdots & \dfrac{\partial p_2}{\partial x_m} & (-p_1 - \hat{x}) \\ \vdots & & \vdots & \vdots \\ \dfrac{\partial p_n}{\partial x_1} & \cdots & \dfrac{\partial p_n}{\partial x_m} & (-p_{n-1} - \hat{x}\, p_{n-2} - \cdots - \hat{x}^{n-2} p_1 - \hat{x}^{n-1}) \\ 0 & \cdots & 0 & (-p_n - \hat{x}\, p_{n-1} - \cdots - \hat{x}^{n-1} p_1 - \hat{x}^n) \end{bmatrix},$$

which has the same rank as J_{X_B}. The leading principal $n \times n$ submatrix of J has full rank. Furthermore, if we substitute the original values corresponding to the entries of A, the $(n+1, n+1)$-entry is a degree-n real polynomial in $\hat{x}$. Thus, there must exist $b > 0$ such that, if $\hat{x} = -b$, this entry is nonzero. Therefore, J has rank $n+1$ after this evaluation. Since $\mathrm{i}(A) = (n_+, n_-, n_0)$, it follows that

$$B = \left[\begin{array}{ccc|c} & & & 0 \\ & A & & \vdots \\ & & & 0 \\ \hline 0 & \cdots & 0 & -b \end{array}\right]$$

is a realization of $\mathcal{A}^-$ that allows a Jacobian of full rank and $\mathrm{i}(B) = (n_+, 1+n_-, n_0)$. □

Combining this result with Theorem 1.1, we obtain the following:

Corollary 3.2. *Let $\mathcal{A}$ and $\mathcal{A}^-$ be defined as above. If $\mathcal{A}$ allows $\mathbb{S}_n$ and has a realization A that allows a Jacobian of full rank and $\mathrm{i}(A) = (0, n-1, 1)$, then every superpattern of $\mathcal{A}^-$ (including $\mathcal{A}^-$ itself) allows $\mathbb{S}_{n+1}$.*

Although $\mathcal{A}^-$ is a reducible matrix, adding at least one additional nonzero entry in the last row and last column of $\mathcal{A}^-$ will yield irreducible patterns of order $n+1$ that allow $\mathbb{S}_{n+1}$. We may repeatedly apply this construction to any matrix that allows $\mathbb{S}_m$ to create large families of irreducible sign patterns that allow $\mathbb{S}_n$ for $n > m$. Below is an example of a family created in such a way. In particular, all $\mathbb{S}_3$-minimal sign patterns were found in Section 2, and all such patterns have a realization A that allows a Jacobian of full rank for which $\mathrm{i}(A) = (0, 2, 1)$. Thus, many families may be created using these patterns as the starting point.

Example 3.3. Using the notation and results of [Berliner et al. 2018], the zero-nonzero pattern

$$\begin{bmatrix} * & * & 0 & * \\ 0 & 0 & * & 0 \\ 0 & * & 0 & * \\ * & 0 & 0 & 0 \end{bmatrix},$$

which corresponds to the digraph G16 with a loop at vertex 1, allows $\mathbb{S}_4$. In fact, there is a corresponding sign pattern $\mathcal{A}$ that allows $\mathbb{S}_4$. The sign pattern $\mathcal{A}$ and a realization $A \in Q(\mathcal{A})$ are given by

$$\mathcal{A} = \begin{bmatrix} - & + & 0 & - \\ 0 & 0 & + & 0 \\ 0 & - & 0 & + \\ + & 0 & 0 & 0 \end{bmatrix} \quad \text{and} \quad A = \begin{bmatrix} -1 & 1 & 0 & -a \\ 0 & 0 & 1 & 0 \\ 0 & -b & 0 & 1 \\ c & 0 & 0 & 0 \end{bmatrix},$$

with $a, b, c > 0$. Without loss of generality, one diagonal entry and two other entries in the digraph associated with $\mathcal{A}$ (corresponding to a spanning tree) can be set equal to ± 1. The characteristic polynomial of A is

$$c_A(z) = z^4 + z^3 + (b + ac)z^2 + bz + (abc - c).$$

In order to realize inertia $(0, 3, 1)$, we must have $c = abc$. If $a = b = c = 1$, then $\mathrm{i}(A) = (0, 3, 1)$, as desired. We now check if A allows a nonzero Jacobian. We replace the nonzero entries of A with variables to get

$$X_A = \begin{bmatrix} x_1 & x_2 & 0 & x_3 \\ 0 & 0 & x_4 & 0 \\ 0 & x_5 & 0 & x_6 \\ x_7 & 0 & 0 & 0 \end{bmatrix},$$

which has characteristic polynomial

$$c_{X_A}(z) = z^4 - x_1 z^3 - (x_3 x_7 + x_4 x_5)z^2 + (x_1 x_4 x_5)z - (x_2 x_4 x_6 x_7 + x_3 x_4 x_5 x_7).$$

Calculating the Jacobian matrix of X_A, we get

$$J_{X_A} = \begin{bmatrix} -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -x_7 & -x_5 & -x_4 & 0 & -x_3 \\ x_4 x_5 & 0 & 0 & x_1 x_5 & x_1 x_4 & 0 & 0 \\ 0 & -x_4 x_6 x_7 & x_4 x_5 x_7 & x_3 x_5 x_7 - x_2 x_6 x_7 & x_3 x_4 x_7 & -x_2 x_4 x_7 & x_3 x_4 x_5 - x_2 x_4 x_6 \end{bmatrix},$$

which when evaluated at $x_1 = x_3 = x_5 = -1,\ x_2 = x_4 = x_6 = x_7 = 1$ has full rank. By Corollary 3.2, any $n \times n$ sign pattern ($n \geq 4$) of the form

$$\begin{bmatrix} - & + & 0 & - & \pm & \cdots & \cdots & \pm \\ 0 & 0 & + & 0 & \pm & & & \vdots \\ 0 & - & 0 & + & \pm & & & \vdots \\ + & 0 & 0 & 0 & \pm & & & \vdots \\ \pm & \pm & \pm & \pm & - & \ddots & & \vdots \\ \vdots & & & & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & & & \ddots & \ddots & \pm \\ \pm & \cdots & \cdots & \cdots & \cdots & \cdots & \pm & - \end{bmatrix}$$

allows $\mathbb{S}_n$ (where the $\pm$ entries may be any of $+$, $-$, or 0).

Appendix A: Nonequivalent superpatterns of $\mathbb{S}_3$-minimal sign patterns

The following sign patterns are the nonequivalent superpatterns of the $\mathbb{S}_3$-minimal sign patterns in Figure 2. All sign patterns with associated digraph D1 that allow

$\mathbb{S}_3$ are $\mathbb{S}_3$-minimal. Thus, the patterns here are organized into four groups corresponding to their associated (loopless) digraph D2–D5. The use of the symbol $\pm$ indicates that a particular entry could be either $-$ or $+$.

Sign patterns with associated digraph D2:

$$\begin{bmatrix} - & - & 0 \\ - & - & - \\ 0 & + & - \end{bmatrix} \begin{bmatrix} - & + & 0 \\ - & + & + \\ 0 & + & - \end{bmatrix} \begin{bmatrix} - & + & 0 \\ - & - & + \\ 0 & - & + \end{bmatrix} \begin{bmatrix} + & - & 0 \\ + & + & - \\ 0 & + & - \end{bmatrix}$$

Sign patterns with associated digraph D3:

$$\begin{bmatrix} - & + & 0 \\ 0 & 0 & + \\ + & - & \pm \end{bmatrix} \begin{bmatrix} - & + & 0 \\ 0 & - & + \\ \pm & + & - \end{bmatrix} \begin{bmatrix} - & + & 0 \\ 0 & \pm & + \\ + & - & - \end{bmatrix} \begin{bmatrix} - & + & 0 \\ 0 & + & + \\ + & - & + \end{bmatrix} \begin{bmatrix} + & + & 0 \\ 0 & - & - \\ + & + & \pm \end{bmatrix} \begin{bmatrix} - & - & 0 \\ 0 & + & + \\ + & + & - \end{bmatrix} \begin{bmatrix} - & + & 0 \\ 0 & + & - \\ + & + & - \end{bmatrix} \begin{bmatrix} + & - & 0 \\ 0 & - & + \\ + & + & - \end{bmatrix}$$

Sign patterns with associated digraph D4:

$$\begin{bmatrix} - & + & 0 \\ + & 0 & + \\ + & - & 0 \end{bmatrix} \begin{bmatrix} - & + & 0 \\ - & 0 & - \\ - & + & 0 \end{bmatrix} \begin{bmatrix} - & - & 0 \\ + & 0 & + \\ + & + & \pm \end{bmatrix} \begin{bmatrix} - & + & 0 \\ - & 0 & + \\ + & + & - \end{bmatrix} \begin{bmatrix} - & + & 0 \\ + & 0 & + \\ - & + & - \end{bmatrix} \begin{bmatrix} - & \pm & 0 \\ - & 0 & + \\ - & - & + \end{bmatrix} \begin{bmatrix} - & + & 0 \\ - & 0 & - \\ - & + & \pm \end{bmatrix} \begin{bmatrix} - & + & 0 \\ + & \pm & + \\ + & - & 0 \end{bmatrix}$$

$$\begin{bmatrix} - & + & 0 \\ + & - & + \\ - & + & 0 \end{bmatrix} \begin{bmatrix} - & + & 0 \\ - & \pm & - \\ - & + & 0 \end{bmatrix} \begin{bmatrix} - & - & 0 \\ + & \pm & + \\ + & + & 0 \end{bmatrix} \begin{bmatrix} + & + & 0 \\ + & - & - \\ + & + & 0 \end{bmatrix} \begin{bmatrix} + & + & 0 \\ - & - & + \\ - & - & 0 \end{bmatrix} \begin{bmatrix} - & \pm & 0 \\ + & + & - \\ + & + & - \end{bmatrix} \begin{bmatrix} - & + & 0 \\ - & \pm & + \\ + & + & - \end{bmatrix} \begin{bmatrix} - & + & 0 \\ + & \pm & + \\ - & + & - \end{bmatrix}$$

$$\begin{bmatrix} - & + & 0 \\ - & - & - \\ - & + & - \end{bmatrix} \begin{bmatrix} + & - & 0 \\ + & - & + \\ + & \pm & + \end{bmatrix} \begin{bmatrix} + & + & 0 \\ - & - & + \\ + & + & + \end{bmatrix} \begin{bmatrix} + & + & 0 \\ + & + & - \\ + & + & - \end{bmatrix} \begin{bmatrix} + & + & 0 \\ - & + & + \\ + & + & - \end{bmatrix} \begin{bmatrix} \pm & - & 0 \\ + & + & - \\ + & + & - \end{bmatrix} \begin{bmatrix} + & - & 0 \\ + & + & + \\ + & - & - \end{bmatrix} \begin{bmatrix} - & - & 0 \\ + & + & + \\ + & \pm & - \end{bmatrix}$$

$$\begin{bmatrix} - & - & 0 \\ + & - & \pm \\ + & + & + \end{bmatrix} \begin{bmatrix} - & + & 0 \\ + & - & - \\ + & + & + \end{bmatrix} \begin{bmatrix} - & \pm & 0 \\ + & - & + \\ + & - & + \end{bmatrix} \begin{bmatrix} - & + & 0 \\ + & - & + \\ - & + & + \end{bmatrix}$$

Sign patterns with associated digraph D5:

$$\begin{bmatrix} - & - & \pm \\ + & 0 & + \\ + & + & 0 \end{bmatrix} \begin{bmatrix} - & \pm & + \\ + & 0 & - \\ + & + & 0 \end{bmatrix} \begin{bmatrix} - & - & - \\ + & 0 & - \\ + & + & 0 \end{bmatrix} \begin{bmatrix} - & - & + \\ + & 0 & \pm \\ + & + & - \end{bmatrix} \begin{bmatrix} - & \pm & - \\ + & 0 & + \\ + & + & - \end{bmatrix} \begin{bmatrix} - & - & + \\ - & 0 & + \\ + & + & - \end{bmatrix}$$

$$\begin{bmatrix} - & - & + \\ + & 0 & + \\ - & + & - \end{bmatrix} \begin{bmatrix} - & - & - \\ + & 0 & - \\ + & + & \pm \end{bmatrix} \begin{bmatrix} - & - & \pm \\ + & 0 & + \\ + & + & + \end{bmatrix} \begin{bmatrix} - & + & - \\ + & 0 & + \\ + & \pm & + \end{bmatrix} \begin{bmatrix} - & \pm & + \\ + & 0 & - \\ + & + & + \end{bmatrix} \begin{bmatrix} - & - & + \\ + & 0 & + \\ + & - & + \end{bmatrix}$$

$$\begin{bmatrix} - & \pm & + \\ + & - & + \\ + & + & - \end{bmatrix} \begin{bmatrix} - & - & - \\ + & - & \pm \\ + & + & - \end{bmatrix} \begin{bmatrix} \pm & - & + \\ - & - & + \\ + & + & - \end{bmatrix} \begin{bmatrix} - & - & + \\ + & - & - \\ + & + & - \end{bmatrix} \begin{bmatrix} - & - & \pm \\ + & + & + \\ + & + & + \end{bmatrix} \begin{bmatrix} - & \pm & + \\ + & + & - \\ + & + & + \end{bmatrix}$$

$$\begin{bmatrix} - & - & + \\ + & + & + \\ + & - & + \end{bmatrix} \begin{bmatrix} - & - & - \\ + & + & - \\ + & + & + \end{bmatrix} \begin{bmatrix} + & - & \pm \\ + & - & + \\ + & + & - \end{bmatrix} \begin{bmatrix} + & \pm & + \\ + & - & - \\ + & + & - \end{bmatrix} \begin{bmatrix} + & - & + \\ + & - & + \\ - & + & - \end{bmatrix} \begin{bmatrix} + & - & + \\ + & - & + \\ + & - & - \end{bmatrix} \begin{bmatrix} + & - & - \\ + & - & - \\ + & + & - \end{bmatrix}$$

Appendix B: Nonequivalent sign-nonsingular sign patterns

Here, we list all nonequivalent sign patterns of order 3 that do not allow $\mathbb{S}_3$. All of these patterns do not allow inertia $(0, 2, 1)$.

B.1. ***Sign-nonsingular sign patterns.*** The following are the nonequivalent sign-nonsingular sign patterns of order 3. Since all realizations of these patterns must be invertible, these patterns do not allow $\mathbb{S}_3$ as their inertias cannot include $(0, 2, 1)$.

The patterns here are organized into five groups corresponding to their associated (loopless) digraph D1–D5.

Sign patterns with associated digraph D1:

$$\begin{bmatrix} - & - & 0 \\ 0 & - & + \\ + & 0 & - \end{bmatrix} \begin{bmatrix} - & - & 0 \\ 0 & + & - \\ - & 0 & + \end{bmatrix} \begin{bmatrix} + & + & 0 \\ 0 & - & + \\ + & 0 & - \end{bmatrix}$$

Sign patterns with associated digraph D2:

$$\begin{bmatrix} - & - & 0 \\ - & 0 & - \\ 0 & - & - \end{bmatrix} \begin{bmatrix} - & + & 0 \\ - & 0 & + \\ 0 & - & - \end{bmatrix} \begin{bmatrix} - & + & 0 \\ - & 0 & - \\ 0 & - & + \end{bmatrix} \begin{bmatrix} - & - & 0 \\ - & 0 & + \\ 0 & - & + \end{bmatrix} \begin{bmatrix} + & + & 0 \\ + & - & + \\ 0 & + & + \end{bmatrix}$$

$$\begin{bmatrix} - & - & 0 \\ - & + & - \\ 0 & - & - \end{bmatrix} \begin{bmatrix} + & + & 0 \\ - & + & + \\ 0 & + & - \end{bmatrix} \begin{bmatrix} - & - & 0 \\ + & - & - \\ 0 & - & + \end{bmatrix} \begin{bmatrix} - & - & 0 \\ + & - & - \\ 0 & + & - \end{bmatrix}$$

Sign patterns with associated digraph D3:

$$\begin{bmatrix} - & + & 0 \\ 0 & 0 & \pm \\ + & + & 0 \end{bmatrix} \begin{bmatrix} - & + & 0 \\ 0 & 0 & \pm \\ + & + & - \end{bmatrix} \begin{bmatrix} + & - & 0 \\ 0 & 0 & + \\ + & + & - \end{bmatrix} \begin{bmatrix} + & + & 0 \\ 0 & 0 & + \\ + & - & - \end{bmatrix} \begin{bmatrix} - & + & 0 \\ 0 & 0 & \pm \\ + & + & + \end{bmatrix}$$

$$\begin{bmatrix} - & + & 0 \\ 0 & - & - \\ + & + & - \end{bmatrix} \begin{bmatrix} - & + & 0 \\ 0 & + & - \\ + & + & + \end{bmatrix} \begin{bmatrix} + & + & 0 \\ 0 & - & + \\ - & + & + \end{bmatrix} \begin{bmatrix} - & + & 0 \\ 0 & + & + \\ + & + & - \end{bmatrix} \begin{bmatrix} + & + & 0 \\ 0 & - & + \\ + & - & - \end{bmatrix}$$

Sign patterns with associated digraph D4:

$$\begin{bmatrix} - & + & 0 \\ + & 0 & \pm \\ + & + & 0 \end{bmatrix} \begin{bmatrix} - & + & 0 \\ - & 0 & + \\ + & + & 0 \end{bmatrix} \begin{bmatrix} - & + & 0 \\ - & 0 & + \\ - & - & 0 \end{bmatrix} \begin{bmatrix} - & + & 0 \\ + & 0 & + \\ + & + & - \end{bmatrix} \begin{bmatrix} - & + & 0 \\ - & 0 & + \\ - & - & - \end{bmatrix} \begin{bmatrix} - & + & 0 \\ - & 0 & + \\ + & + & + \end{bmatrix} \begin{bmatrix} - & - & 0 \\ - & 0 & - \\ - & + & + \end{bmatrix} \begin{bmatrix} - & + & 0 \\ + & - & \pm \\ + & + & 0 \end{bmatrix}$$

$$\begin{bmatrix} - & + & 0 \\ - & \pm & + \\ + & + & 0 \end{bmatrix} \begin{bmatrix} - & + & 0 \\ - & \pm & + \\ - & - & 0 \end{bmatrix} \begin{bmatrix} - & + & 0 \\ + & + & \pm \\ + & + & 0 \end{bmatrix} \begin{bmatrix} + & - & 0 \\ + & - & + \\ + & + & 0 \end{bmatrix} \begin{bmatrix} + & + & 0 \\ + & - & + \\ + & - & 0 \end{bmatrix} \begin{bmatrix} + & + & 0 \\ + & - & + \\ - & + & 0 \end{bmatrix} \begin{bmatrix} + & + & 0 \\ - & - & - \\ - & + & 0 \end{bmatrix} \begin{bmatrix} - & + & 0 \\ - & - & + \\ - & - & - \end{bmatrix}$$

$$\begin{bmatrix} + & + & 0 \\ + & - & + \\ - & + & + \end{bmatrix} \begin{bmatrix} + & - & 0 \\ + & + & + \\ + & + & - \end{bmatrix} \begin{bmatrix} - & + & 0 \\ + & + & + \\ + & + & - \end{bmatrix} \begin{bmatrix} - & + & 0 \\ - & - & + \\ + & + & + \end{bmatrix}$$

Sign patterns with associated digraph D5:

$$\begin{bmatrix} - & + & + \\ + & 0 & + \\ + & + & 0 \end{bmatrix} \begin{bmatrix} - & - & + \\ + & 0 & + \\ + & - & 0 \end{bmatrix} \begin{bmatrix} - & - & + \\ + & 0 & + \\ - & + & 0 \end{bmatrix} \begin{bmatrix} - & + & + \\ + & 0 & + \\ + & + & - \end{bmatrix} \begin{bmatrix} - & - & + \\ + & 0 & + \\ + & - & - \end{bmatrix} \begin{bmatrix} - & - & + \\ + & 0 & + \\ - & + & + \end{bmatrix} \begin{bmatrix} - & + & - \\ + & 0 & - \\ + & + & + \end{bmatrix}$$

B.2. ***Other sign patterns that do not allow*** **(0, 2, 1).** The following sign patterns are not sign-nonsingular, but nevertheless do not allow inertia $(0, 2, 1)$. In the characteristic polynomial of a realization, it can be shown that it is impossible for the constant term to equal 0 when all of the other coefficients are positive. The patterns are organized into five groups corresponding to their associated (loopless) digraph D1–D5.

Sign pattern with associated digraph D1:

$$\begin{bmatrix} - & + & 0 \\ 0 & + & + \\ + & 0 & + \end{bmatrix}$$

Sign patterns with associated digraph D2:

$$\begin{bmatrix} - & - & 0 \\ - & 0 & - \\ 0 & - & + \end{bmatrix} \begin{bmatrix} + & + & 0 \\ + & + & + \\ 0 & \pm & - \end{bmatrix} \begin{bmatrix} - & - & 0 \\ - & - & - \\ 0 & - & + \end{bmatrix} \begin{bmatrix} + & + & 0 \\ - & - & + \\ 0 & - & + \end{bmatrix}$$

Sign patterns with associated digraph D3:

$$\begin{bmatrix} - & - & 0 \\ 0 & 0 & + \\ + & + & 0 \end{bmatrix} \begin{bmatrix} + & + & 0 \\ 0 & 0 & + \\ + & + & - \end{bmatrix} \begin{bmatrix} - & - & 0 \\ 0 & 0 & + \\ + & + & + \end{bmatrix} \begin{bmatrix} - & + & 0 \\ 0 & + & + \\ \pm & + & + \end{bmatrix} \begin{bmatrix} + & + & 0 \\ 0 & - & + \\ + & \pm & + \end{bmatrix} \begin{bmatrix} + & + & 0 \\ 0 & - & + \\ + & + & - \end{bmatrix}$$

Sign patterns with associated digraph D4:

$$\begin{bmatrix} - & + & 0 \\ + & 0 & + \\ - & + & 0 \end{bmatrix} \begin{bmatrix} - & + & 0 \\ + & 0 & + \\ \pm & + & + \end{bmatrix} \begin{bmatrix} - & + & 0 \\ + & + & + \\ - & + & 0 \end{bmatrix} \begin{bmatrix} + & + & 0 \\ + & - & + \\ + & + & 0 \end{bmatrix} \begin{bmatrix} \pm & + & 0 \\ + & - & + \\ + & + & + \end{bmatrix} \begin{bmatrix} + & - & 0 \\ + & - & - \\ + & + & + \end{bmatrix} \begin{bmatrix} + & + & 0 \\ + & + & + \\ + & \pm & - \end{bmatrix} \begin{bmatrix} + & + & 0 \\ + & + & + \\ - & + & - \end{bmatrix}$$

Sign patterns with associated digraph D5:

$$\begin{bmatrix} - & - & + \\ - & 0 & + \\ + & + & 0 \end{bmatrix} \begin{bmatrix} - & + & + \\ + & 0 & + \\ + & + & + \end{bmatrix} \begin{bmatrix} - & - & + \\ - & 0 & + \\ + & + & + \end{bmatrix} \begin{bmatrix} - & + & + \\ + & + & + \\ + & + & + \end{bmatrix} \begin{bmatrix} - & - & + \\ - & + & + \\ \pm & + & + \end{bmatrix} \begin{bmatrix} + & + & + \\ + & - & + \\ + & + & - \end{bmatrix}$$

Acknowledgements

The authors would like to thank the St. Olaf College MSCS Department and CURI Program for their support of this research.

References

[Berliner et al. 2017] A. H. Berliner, D. D. Olesky, and P. van den Driessche, "Sets of refined inertias of zero-nonzero patterns", *Linear Algebra Appl.* **516** (2017), 243–263. MR Zbl

[Berliner et al. 2018] A. H. Berliner, D. D. Olesky, and P. van den Driessche, "Inertia sets allowed by matrix patterns", *Electron. J. Linear Algebra* **34** (2018), 343–355. MR Zbl

[Bodine et al. 2012] E. Bodine, L. Deaett, J. J. McDonald, D. D. Olesky, and P. van den Driessche, "Sign patterns that require or allow particular refined inertias", *Linear Algebra Appl.* **437**:9 (2012), 2228–2242. MR Zbl

[Cavers and Vander Meulen 2005] M. S. Cavers and K. N. Vander Meulen, "Spectrally and inertially arbitrary sign patterns", *Linear Algebra Appl.* **394** (2005), 53–72. MR Zbl

[Gao et al. 2016a] W. Gao, Z. Li, and L. Zhang, "Characterization of star sign patterns that require $\mathbb{H}_n$", *Linear Algebra Appl.* **499** (2016), 43–65. MR Zbl

[Gao et al. 2016b] W. Gao, Z. Li, and L. Zhang, "Sign patterns that require $\mathbb{H}_n$ exist for each $n \geq 4$", *Linear Algebra Appl.* **489** (2016), 15–23. MR Zbl

[Garnett and Shader 2013] C. Garnett and B. L. Shader, "The nilpotent-centralizer method for spectrally arbitrary patterns", *Linear Algebra Appl.* **438**:10 (2013), 3836–3850. MR Zbl

[Garnett et al. 2013] C. Garnett, D. D. Olesky, and P. van den Driessche, "Refined inertias of tree sign patterns", *Electron. J. Linear Algebra* **26** (2013), 620–635. MR Zbl

[Garnett et al. 2014] C. Garnett, D. D. Olesky, and P. van den Driessche, "A note on sign patterns of order 3 that require particular refined inertias", *Linear Algebra Appl.* **450** (2014), 293–300. MR Zbl

[Olesky et al. 2013] D. D. Olesky, M. F. Rempel, and P. van den Driessche, "Refined inertias of tree sign patterns of orders 2 and 3", *Involve* **6**:1 (2013), 1–12. MR Zbl

Received: 2019-01-29 Revised: 2019-06-08 Accepted: 2019-06-22

berliner@stolaf.edu *Department of Mathematics, Statistics, and Computer Science, St. Olaf College, Northfield, MN, United States*

derek.deblieck@gmail.com *St. Olaf College, Northfield, MN, United States*

kalopatthar@gmail.com *St. Olaf College, Northfield, MN, United States*

INVOLVE 12:7 (2019)
dx.doi.org/10.2140/involve.2019.12.1241

Some combinatorics from Zeckendorf representations

Tyler Ball, Rachel Chaiser, Dean Dustin, Tom Edgar and Paul Lagarde

(Communicated by Arthur T. Benjamin)

We explore some properties of the so-called Zeckendorf representations of integers, where we write an integer as a sum of distinct, nonconsecutive Fibonacci numbers. We examine the combinatorics arising from the arithmetic of these representations, with a particular emphasis on understanding the Zeckendorf tree that encodes them. We introduce some possibly new results related to the tree, allowing us to develop a partial analog to Kummer's classical theorem about counting the number of "carries" involved in arithmetic. Finally, we finish with some conjectures and possible future projects related to the combinatorics of these representations.

1. Introduction

Given an integer $b \geq 2$, we can write each natural number uniquely in its base-b representation $n = \sum_{i=0}^{k} n_i b^i$, where $0 \leq n_i < b$ and $n_k \neq 0$. The classical version of Kummer's theorem yields a connection between the prime factorizations of binomial coefficients and base-p arithmetic.

Theorem 1.1 (Kummer). *Let $n, m, p \in \mathbb{N}$ with p prime. Then the exponent of the largest power of p dividing $\binom{n+m}{n}$ is the sum of the carries when adding the base-p representations of n and m.*

Ball et al. [2014] constructed families of generalized binomial coefficients demonstrating a similar phenomenon for base-b arithmetic even when b is composite, and Edgar et al. [2014] did the same for rational base arithmetic, i.e., base-p/q when $p > q \geq 1$ are relatively prime integers.

In this paper, we consider the so-called Zeckendorf representation of integers, where we write an integer as the sum of distinct, nonconsecutive Fibonacci numbers, and we construct a family of generalized binomial coefficients that provide partial information about the "carries" involved in the arithmetic of Zeckendorf representations. The paper is organized as follows. In Section 2, we formally

MSC2010: 06A07, 11B39, 11B75, 11Y55.
Keywords: Fibonacci, Zeckendorf, digital dominance order.

introduce Zeckendorf representations, some related combinatorial structures, and some relevant integer sequences. In particular, we include some results that are likely already known but for which proofs and citations are difficult to find. In Section 3, we describe one method for adding Zeckendorf representations and then discuss our main result; we construct a sequence whose generalized binomial coefficients give us the appropriate generalization of Kummer's theorem for Zeckendorf representations. Finally, in Section 4, we discuss how these results are related to a partial order on the set of natural numbers arising from Zeckendorf representations and describe some questions and conjectures arising from this partial order and Zeckendorf arithmetic.

2. Preliminaries and background

Let $F : \mathbb{N} \to \mathbb{N}$ be the Fibonacci sequence defined by $F(0) = 0$, $F(1) = 1$, and $F(n) = F(n-1) + F(n-2)$ for $n \geq 2$, and let $f : \mathbb{N} \to \mathbb{N}$ be the related sequence defined by $f(n) = F(n+1)$; we will write f_i in place of $f(i)$. The sequence f yields the standard indexing for a combinatorial interpretation of the Fibonacci sequence, as f_n counts the number of ways to tile an $n \times 1$ board with tiles of size 1×1 and 1×2.

Now, a Fibonacci representation of a natural number n is a list $(n_1, n_2, \ldots, n_k)_f$, satisfying $n = \sum_{i=1}^{k} n_i f_i$, $n_k \neq 0$, and $n_i \in \{0, 1\}$. Unfortunately, Fibonacci representations are not necessarily unique for a given natural number. For example, the number 6 has exactly two Fibonacci representations:

$$6 = 1\cdot 1 + 1\cdot 2 + 1\cdot 3 = (1, 1, 1)_f \quad \text{and} \quad 6 = 1\cdot 1 + 0\cdot 2 + 0\cdot 3 + 1\cdot 5 = (1, 0, 0, 1)_f.$$

To guarantee uniqueness, we add an extra condition: we require that there are never consecutive 1's in the list. More formally, the *Zeckendorf representation* of n is the list $(n_1, n_2, \ldots, n_k)_z$, where $n = \sum_{i=1}^{k} n_i f_i$, $n_k \neq 0$, $n_i \in \{0, 1\}$, and $n_i \cdot n_{i+1} = 0$ for all $i < k$. It is well known that the Zeckendorf representation is unique [Zeckendorf 1972]. We will write $n = (n_1, n_2, \ldots, n_k)_z$ to mean that $(n_1, n_2, \ldots, n_k)_z$ is the Zeckendorf representation of n; note that the list is written in order from the least-significant to most-significant digit. We may often also refer to n_i when $i > k$, in which case we mean $n_i = 0$ since appending 0's to the list will not change the value of the sum.

Next, we can define $s_z(n) = \sum_{i=1}^{k} n_i$ when $n = (n_1, n_2, \ldots, n_k)_z$; we call s_z the *Zeckendorf sum-of-digits function*. For instance, the following lemma determines the Zeckendorf sum-of-digits for numbers that are one less than a Fibonacci number (note that $s_z(f_i) = 1$ for all $i \geq 1$).

Lemma 2.1. *Let $\ell \in \mathbb{N}$. Then*

$$f_{2\ell} - 1 = (\underbrace{1, 0, 1, 0, \ldots, 1}_{2\ell-1})_z, \qquad f_{2\ell+1} - 1 = (\underbrace{0, 1, 0, 1, 0, \ldots, 1}_{2\ell})_z.$$

Consequently, $s_z(f_j - 1) = \left\lfloor \frac{1}{2} j \right\rfloor$ for all j.

Proof. These are standard Fibonacci identities and can be found, for instance, in [Benjamin and Quinn 2003]. □

Many sources (for instance [Marsault and Sakarovitch 2014; 2017] and the Sillke link from A005206 on [OEIS]) describe a tree structure for building the Zeckendorf representations but don't include proof. We describe two functions that allow us to build this Zeckendorf representation tree that make it clear the tree does in fact give the Zeckendorf representations. In particular, we define $b:\mathbb{N}\to\mathbb{N}$ and $p:\mathbb{N}\to\mathbb{N}$ by

$$b(n)=\lfloor(n+1)\phi\rfloor-1,\quad p(n)=\left\lfloor\frac{n+2}{\phi}\right\rfloor-1,$$

where $\phi=\frac{1}{2}(1+\sqrt{5})$ is the golden ratio (i.e., the unique positive solution to the equation $x^2-x-1=0$).

The function b seems to have been widely studied (for instance see [Kimberling 1995]) and is given by A022342 in [OEIS]. The function p is missing from OEIS; however, after some searching we did discover that $p(n)+1$ is A005206 in [OEIS] and contains a link due to Sillke that mentions its connection (without proof) to Zeckendorf representations. The following (seemingly known) theorem describes the relevance of the two functions to Zeckendorf representations.

Theorem 2.2. *Let $n=(n_1,n_2,n_3,\ldots,n_k)_z$. Then $b(n)=(0,n_1,n_2,n_3,\ldots,n_k)_z$ and $p(n)=(n_2,n_3,\ldots,n_k)_z$.*

Proof. Let $n=(n_1,n_2,n_3,\ldots,n_k)_z$ so that $n=\sum_{i=1}^k n_i f_i$. Using the Binet formula for Fibonacci numbers, which says $f_i=(\phi^{i+1}-\phi^{-(i+1)})/\sqrt{5}$, we have

$$\begin{aligned}
b(n)=\lfloor(n+1)\phi\rfloor-1&=\left\lfloor\left(\sum_{i=1}^k n_i\left(\frac{\phi^{i+1}-\phi^{-(i+1)}}{\sqrt{5}}\right)\phi\right)+\phi\right\rfloor-1\\
&=\left\lfloor\left(\sum_{i=1}^k n_i\left(\frac{\phi^{i+2}-\phi^{-i}}{\sqrt{5}}\right)\right)+\phi\right\rfloor-1\\
&=\left\lfloor\left(\sum_{i=1}^k n_i\left(\frac{\phi^{i+2}-\phi^{-(i+2)}}{\sqrt{5}}+\frac{\phi^{-(i+2)}-\phi^{-i}}{\sqrt{5}}\right)\right)+\phi\right\rfloor-1\\
&=\left\lfloor\sum_{i=1}^k n_i\left(\frac{\phi^{i+2}-\phi^{-(i+2)}}{\sqrt{5}}\right)+\sum_{i=1}^k n_i\left(\frac{\phi^{-(i+2)}-\phi^{-i}}{\sqrt{5}}\right)+\phi\right\rfloor-1\\
&=\left\lfloor\sum_{i=1}^k n_i f_{i+1}+\sum_{i=1}^k n_i\left(\frac{-\phi^{-(i+1)}}{\sqrt{5}}\right)+\phi\right\rfloor-1\\
&=\sum_{i=1}^k n_i f_{i+1}+\left\lfloor\sum_{i=1}^k n_i\left(\frac{-\phi^{-(i+1)}}{\sqrt{5}}\right)+\phi\right\rfloor-1.
\end{aligned}$$

The second-to-last inequality uses the Binet formula and that $\phi^{-(i+2)}-\phi^{-i}=\phi^{-(i+1)}$.

Thus, it suffices to demonstrate that

$$1 \leq \sum_{i=1}^{k} n_i \left(\frac{-\phi^{-(i+1)}}{\sqrt{5}} \right) + \phi < 2.$$

First, we see that $\phi - \sum_{i=1}^{k} n_i (\phi^{-(i+1)}/\sqrt{5}) \leq \phi < 2$. Next, we note that $0 \leq n_i \leq 1$ for all i so that

$$\phi - \sum_{i=1}^{\infty} \left(\frac{\phi^{-(i+1)}}{\sqrt{5}} \right) \leq \phi - \sum_{i=1}^{k} \left(\frac{\phi^{-(i+1)}}{\sqrt{5}} \right) \leq \phi - \sum_{i=1}^{k} n_i \left(\frac{\phi^{-(i+1)}}{\sqrt{5}} \right).$$

However, the series $\sum_{i=1}^{\infty} (\phi^{-(i+1)}/\sqrt{5})$ is a geometric series so that

$$\sum_{i=1}^{\infty} \left(\frac{\phi^{-(i+1)}}{\sqrt{5}} \right) = \frac{\phi^{-2}}{\sqrt{5}(1-\phi^{-1})} = \frac{1}{\sqrt{5}},$$

which means

$$1 < \phi - \frac{1}{\sqrt{5}} \leq \phi - \sum_{i=1}^{k} n_i \left(\frac{\phi^{-(i+1)}}{\sqrt{5}} \right),$$

as required. Thus, we see that $b(n) = \sum_{i=1}^{k} n_i f_{i+1}$; i.e., $b(n) = (0, n_1, n_2, \ldots, n_k)$.

For the second part, we again use Binet's formula and see that

$$\begin{aligned}
p(n) &= \left\lfloor \frac{n+2}{\phi} \right\rfloor - 1 = \left\lfloor \frac{1}{\phi} \left(\sum_{i=1}^{k} (n_i f_i) + 2 \right) \right\rfloor - 1 \\
&= \left\lfloor \frac{1}{\phi} \sum_{i=2}^{k} (n_i f_i) + \frac{2+n_1}{\phi} \right\rfloor - 1 \\
&= \left\lfloor \frac{1}{\phi} \sum_{i=2}^{k} \left(n_i \frac{\phi^{i+1} - \phi^{-(i+1)}}{\sqrt{5}} \right) + \frac{2+n_1}{\phi} \right\rfloor - 1 \\
&= \left\lfloor \sum_{i=2}^{k} n_i \left(\frac{\phi^{i} - \phi^{-(i+2)}}{\sqrt{5}} \right) + \frac{2+n_1}{\phi} \right\rfloor - 1 \\
&= \left\lfloor \sum_{i=2}^{k} n_i \left(\frac{\phi^{i} - \phi^{-i} + \phi^{-i} - \phi^{-(i+2)}}{\sqrt{5}} \right) + \frac{2+n_1}{\phi} \right\rfloor - 1 \\
&= \left\lfloor \sum_{i=2}^{k} n_i \left(\frac{\phi^{i} - \phi^{-i}}{\sqrt{5}} + \frac{\phi^{-i} - \phi^{-(i+2)}}{\sqrt{5}} \right) + \frac{2+n_1}{\phi} \right\rfloor - 1 \\
&= \left\lfloor \sum_{i=2}^{k} n_i f_{i-1} + \sum_{i=2}^{k} n_i \left(\frac{\phi^{-i} - \phi^{-(i+2)}}{\sqrt{5}} \right) + \frac{2+n_1}{\phi} \right\rfloor - 1 \\
&= \sum_{i=2}^{k} n_i f_{i-1} + \left\lfloor \sum_{i=2}^{k} n_i \left(\frac{\phi^{-(i+1)}}{\sqrt{5}} \right) + \frac{2+n_1}{\phi} \right\rfloor - 1.
\end{aligned}$$

Thus, again, it will suffice to show that

$$1 \le \sum_{i=2}^{k} n_i \left(\frac{\phi^{-(i+1)}}{\sqrt{5}} \right) + \frac{2+n_1}{\phi} < 2.$$

To do this, we consider two cases: $n_1 = 0$ and $n_1 = 1$.

Case 1: Let $n_1 = 0$. Then $n_{2\ell}(\phi^{-(2\ell+1)}/\sqrt{5}) + n_{2\ell+1}(\phi^{-(2\ell+2)}/\sqrt{5}) \le (\phi^{-(2\ell+1)}/\sqrt{5})$ for all $\ell \ge 1$ since at most one of $n_{2\ell}$ and $n_{2\ell+1}$ is nonzero. This fact along with $n_1 = 0$ implies

$$\begin{aligned} 1 < \frac{2}{\phi} &\le \sum_{i=2}^{k} n_i \left(\frac{\phi^{-(i+1)}}{\sqrt{5}} \right) + \frac{2+n_1}{\phi} \\ &\le \sum_{\ell=1}^{\infty} \left(\frac{\phi^{-(2\ell+1)}}{\sqrt{5}} \right) + \frac{2}{\phi} = \frac{\phi^{-3}}{\sqrt{5}(1-\phi^{-2})} + \frac{2}{\phi} = \frac{1+2\sqrt{5}\phi}{\sqrt{5}\phi^2} < 2, \end{aligned}$$

as required.

Case 2: Let $n_1 = 1$. Then $n_2 = 0$ and $n_{2\ell-1}(\phi^{-2\ell}/\sqrt{5}) + n_{2\ell}(\phi^{-(2\ell+1)}/\sqrt{5}) \le (\phi^{-2\ell}/\sqrt{5})$ for all $\ell \ge 2$ since at most one of $n_{2\ell-1}$ and $n_{2\ell}$ is nonzero. This fact along with $n_1 = 0$ implies

$$\begin{aligned} 1 < \frac{2}{\phi} &\le \sum_{i=2}^{k} n_i \left(\frac{\phi^{-(i+1)}}{\sqrt{5}} \right) + \frac{2+n_1}{\phi} \\ &\le \sum_{\ell=2}^{\infty} \left(\frac{\phi^{-2\ell}}{\sqrt{5}} \right) + \frac{3}{\phi} = \frac{\phi^{-4}}{\sqrt{5}(1-\phi^{-2})} + \frac{3}{\phi} = \frac{1+3\sqrt{5}\phi^2}{\sqrt{5}\phi^3} < 2, \end{aligned}$$

as required.

So, in either case, we see that $p(n) = \sum_{i=2}^{k} n_i f_{i-1}$, as we wanted to show. □

We can also investigate a few properties of these integer sequences.

Corollary 2.3. *Let $n = \sum_{i=1}^{k} n_i f_i$. Then $n_1 = 0$ if and only if $b(p(n)) = n$.*

Proof. By Theorem 2.2, we note that

$$b(p(n)) = b\left(p\left(\sum_{i=1}^{k} n_i f_i \right) \right) = b\left(\sum_{i=2}^{k} n_i f_{i-1} \right) = \sum_{i=2}^{k} n_i f_i,$$

so that $n - b(p(n)) = n_1 f_1 = n_1$. □

The previous proof provides a formula for n_1. We can extend this idea to provide a formula for each digit in the Zeckendorf representation of a number.

Corollary 2.4. *Let $n = \sum_{i=1}^{k} n_i f_i$. Then $n_j = p^{j-1}(n) - b(p^j(n))$ for any $1 \le j \le k$.*

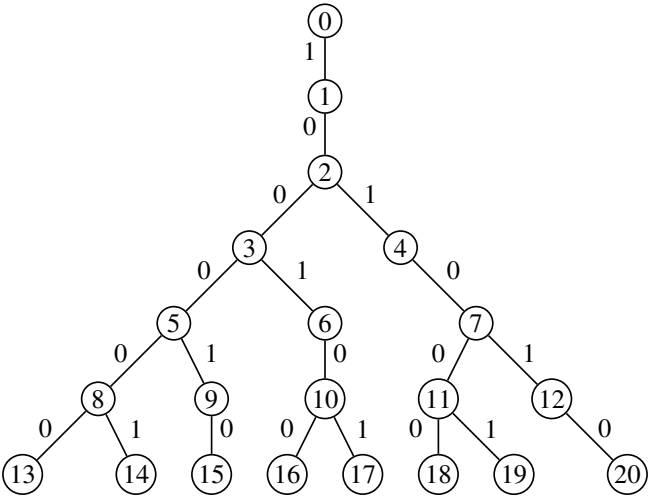

Figure 1. The Zeckendorf tree up to $n = 20$.

Proof. By repeated application of Theorem 2.2, we see

$$\begin{aligned} p^{j-1}(n) - b(p^j(n)) &= \sum_{i=1+j-1}^{k} n_i f_{i-j+1} - b\Big(\sum_{i=1+j}^{k} n_i f_{i-j}\Big) \\ &= \sum_{i=j}^{k} n_i f_{i-j+1} - \sum_{i=1+j}^{k} n_i f_{i-j+1} \\ &= n_j f_1 + \sum_{i=j+1}^{k} n_i f_{i-j+1} - \sum_{i=j+1}^{k} n_i f_{i-j+1} = n_j. \end{aligned}$$

□

A greedy algorithm is typically used to find the Zeckendorf representation of a positive integer, but Corollary 2.4 provides an alternate method for producing the representation, and this method can be encoded in a tree. Consider the graph $(\mathbb{N}, E)$ with vertex set $\mathbb{N}$ and edge set defined by $E = \{\{n, p(n)\} \mid n \in \mathbb{N} \setminus \{0\}\}$. This graph is a tree with root 0. Furthermore, we can define an edge-label function $e : E \to \{0, 1\}$ by $e(n) = n - b(p(n))$. We call this labeled tree the *Zeckendorf tree*, and we have drawn this tree (up to $n = 20$) in Figure 1. In the Zeckendorf tree, we refer to the vertex $p(n)$ as the *parent* of n, the vertex $b(n)$ as the *young child* of n, and the vertex $b(n) + 1$ as the *old child* of n. For example, 10 is the parent of 16 and 17, where 16 is the old child of 10 and 17 is the young child of 10; since 12 has only one child, 20, we refer to 20 as the old child of 12.

As noted, we are not the first to describe this tree, and it can be found various places in the literature. However, this tree is often constructed by the out-degrees of vertices (see [Marsault and Sakarovitch 2014; 2017]); by our construction and

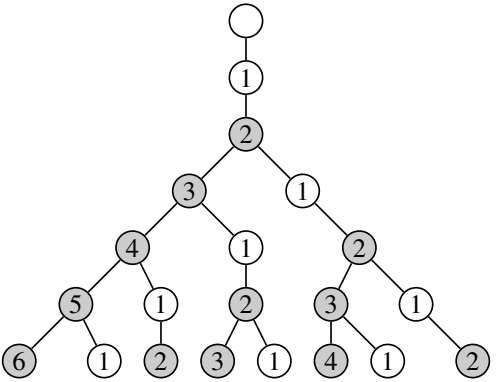

Figure 2. Zeckendorf tree with alternate labels. The shaded vertices represent old children so the labels are thus one more than the parent. The other nodes are young children and hence the labels are 1.

Corollaries 2.3 and 2.4, it is clear that the edge labels on the unique path from n to 0 do in fact yield the Zeckendorf representation of n.

Corollary 2.5. *The list of edge labels on the path from n to 0 in the Zeckendorf tree gives the Zeckendorf representation of n.*

For instance we see that $15 = 0\cdot1+1\cdot2+0\cdot3+0\cdot5+0\cdot8+1\cdot13 = (0, 1, 0, 0, 0, 1)_z$ and these are precisely the edge labels on the path from 15 to 0 in the Zeckendorf tree.

Next, for $n = (n_1, n_2, \ldots, n_k)_z$, we let $w(n) = \min\{i \mid n_i = 1\}$; i.e., $f_{w(n)}$ is the least Fibonacci number used in the Zeckendorf representation of n. Thus, $w(n) - 1$ counts the number of 0's at the beginning of the Zeckendorf representation. The sequence w is given by A035612 in [OEIS] and has been extensively studied since it is connected to the Wythoff array (A035513). The first few values of w are listed in the following table:

n	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
$w(n)$	1	2	3	1	4	1	2	5	1	2	3	1	6	1	2	3	1	4	1	2

Now, consider the following alternative labeling of the vertices of the Zeckendorf tree (see Figure 2):

(1) Omit a label on the vertex 0 and label vertex 1 with 1.

(2) For $n > 1$, if n is a young child, label it 1.

(3) If n is an old child, label it with one more than its parent's label.

In light of Corollary 2.3, we see that this labeling produces the function w.

Theorem 2.6. *For $n \geq 1$,*

$$w(n) = \begin{cases} w(p(n)) + 1 & \text{if } b(p(n)) = n, \\ 1 & \text{otherwise.} \end{cases}$$

We are not aware if this recursive formula for A035612 was previously known. We can also use our results to establish some further results about this and related integer sequences that will be utilized in the following section.

Theorem 2.7. *For $n \geq 1$,*

$$\left\lfloor \tfrac{1}{2} w(n) \right\rfloor = 1 + s_z(n-1) - s_z(n),$$
$$\left\lfloor \tfrac{1}{2}(w(n)-1) \right\rfloor = 1 - n_1 + s_z(p(n-1)) - s_z(p(n)).$$

Proof. First, we note that the Zeckendorf representation of n is $n = \sum_{i=w(n)}^{k} n_i f_i$; in particular, we note that $n = (0, 0, \ldots, 0, 1, n_{w(n)+1}, \ldots, n_k)_z$. Thus, there exists an integer m with $n = f_{w(n)} + m$, so that $m = (0, 0, \ldots, 0, 0, n_{w(n)+1}, \ldots, n_k)_z$, and $s_z(n) = s_z(f_{w(n)}) + s_z(m)$. Also, by Lemma 2.1 we can see that $n-1 = (f_{w(n)}-1)+m$ so that $s_z(n-1) = s_z(f_{w(n)} - 1) + s_z(m)$. We thus have

$$1 + s_z(n-1) - s_z(n) = 1 + s_z(f_{w(n)} - 1) + s_z(m) - (s_z(f_{w(n)}) + s_z(m)) = s_z(f_{w(n)} - 1),$$

since $s_z(f_{w(n)}) = 1$. Thus, again by Lemma 2.1, we have

$$1 + s_z(n-1) - s_z(n) = s_z(f_{w(n)} - 1) = \left\lfloor \tfrac{1}{2} w(n) \right\rfloor,$$

as required for the first part.

The second part can be shown using two cases. First, we suppose that $w(n) \neq 1$. Then $p(n) = (0, \ldots, 0, 1, n_{w(n)+1}, \ldots, n_k)_z$ and so $s_z(p(n)) = 1 + s_z(m)$. Now, again we know that $n - 1 = (f_{w(n)} - 1) + m$, which implies $s_z(p(n-1)) = s_z(f_{w(n)-1} - 1) + s_z(m)$. Putting these together, we see

$$s_z(p(n-1)) - s_z(p(n)) = s_z(f_{w(n)} - 1) + s_z(m) - (1 + s_z(m)) = \left\lfloor \tfrac{1}{2}(w(n)-1) \right\rfloor - 1.$$

Next, we suppose that $w(n) = 1$. Then Theorem 2.2 implies $p(n) = p(n-1)$ so that

$$s_z(p(n-1)) - s_z(p(n)) = 0 = \left\lfloor \tfrac{1}{2}(w(n)-1) \right\rfloor.$$

The result now follows. □

The previous result allows us to give a closed form for w in terms of the sum-of-digits function.

Corollary 2.8. *For $n \geq 1$, $w(n) = 3 - n_1 + s_z(n-1) - s_z(n) + s_z(p(n-1)) - s_z(p(n))$.*

Proof. First, we know that for any natural number a,

$$\left\lfloor \tfrac{1}{2} a \right\rfloor + \left\lfloor \tfrac{1}{2}(a-1) \right\rfloor = a - 1$$

since 2 divides either a or $a-1$ but not both. Thus, by Theorem 2.7

$$\begin{aligned} w(n) &= \left\lfloor \tfrac{1}{2} w(n) \right\rfloor + \left\lfloor \tfrac{1}{2}(w(n)-1) \right\rfloor + 1 \\ &= 1 + s_z(n-1) - s_z(n) + 1 - n_1 + s_z(p(n-1)) - s_z(p(n)) + 1 \\ &= 3 - n_1 + s_z(n-1) - s_z(n) + s_z(p(n-1)) - s_z(p(n)). \end{aligned}$$ □

We will make use of these functions in the next section as we describe our analog of Kummer's theorem.

3. Zeckendorf arithmetic, generalized binomial coefficients and Kummer's theorem

In order to generalize Kummer's theorem to Zeckendorf representations, we will describe one algorithm for producing the Zeckendorf representations of the sum of two numbers using only their Zeckendorf representations; moreover, we will also construct a suitable replacement for binomial coefficients.

Fenwick [2003] demonstrated a method for determining the Zeckendorf representation for $a+b$ in terms of the Zeckendorf representations for a and b. Ahlbach et al. [2013] described a more efficient method of performing the same task based on a result of [Frougny 1991]. Here, we provide a slight modification of the arithmetic described in [Fenwick 2003] instead of the more efficient algorithm due to the fact that this requires us to remember fewer rules. The rules that we employ depend on the fact that the defining relation of the Fibonacci numbers, $f_{n+1} = f_n + f_{n-1}$, implies $2f_n = f_{n+1} + f_{n-2}$ when $n > 2$ (along with the facts that $2f_2 = f_1 + f_3$ and $2f_1 = f_2$). Using these facts, we have the following four rules that we use to transform a list into another list:

Rule 1: $(\ldots, x, y, \mathbf{2}, z, \ldots) \mapsto (\ldots, x+1, y, \mathbf{0}, z+1, \ldots)$.

Rule 2: $(x, \mathbf{2}, y, \ldots) \mapsto (x+1, \mathbf{0}, y+1, \ldots)$.

Rule 3: $(\mathbf{2}, x, \ldots) \mapsto (\mathbf{0}, x+1, \ldots)$.

Rule 4: $(\ldots, 1, \mathbf{1}, x, \ldots) \mapsto (\ldots, 0, \mathbf{0}, x+1, \ldots)$.

Now, given two Zeckendorf representations, $n = (n_1, n_2, \ldots, n_k)_z$ and $m = (m_1, m_2, \ldots, m_k)_z$ (where we can append 0's to ensure both lists are the same length), we can obtain the Zeckendorf representation for $n+m$ by using the following procedure, which has three stages.

Stage 1: Add the two lists $(n_1, n_2, \ldots, n_k)_z$ and $(m_1, m_2, \ldots, m_k)_z$ digit by digit to produce the new list $(n_1+m_1, n_2+m_2, \ldots, n_k+m_k)$.

Stage 2: From left to right (least significant to most significant), apply Rules 1, 2, and 3 until the list contains no 2's.

$$
\begin{aligned}
& & (1,0,1,0,1,0)_z \\
& + & (1,0,1,0,1,0)_z \\
\text{add digits} & \longrightarrow & (2,0,2,0,2,0) \\
\text{Rule 3} & \longrightarrow & (0,1,2,0,2,0) \\
\text{Rule 1} & \longrightarrow & (1,1,0,1,2,0) \\
\text{Rule 1} & \longrightarrow & (1,1,1,1,0,1) \\
\text{Rule 4} & \longrightarrow & (1,1,0,0,1,1) \\
\text{Rule 4} & \longrightarrow & (1,1,0,0,0,0,1) \\
\text{Rule 4} & \longrightarrow & (0,0,1,0,0,0,1)_z
\end{aligned}
$$

Figure 3. Addition of Zeckendorf representations: adding $12+12=24$.

Stage 3: From right to left (most significant to least significant), apply Rule 4 until the list contains no consecutive 1's.

Note that after applying a rule while in Stage 2 or Stage 3, we must start again from the left/right of the list since we might have created an earlier 2 (in Stage 2) or a later instance of (1, 1) (in Stage 3). For example, Figure 3 demonstrates this algorithm when finding the Zeckendorf representation of $24 = (0,0,1,0,0,0,1)_z$ given that $12 = (1,0,1,0,1)_z$ and $12+12=24$.

As noted, this algorithm is far from efficient but will terminate according to [Fenwick 2003].

We refer to each of the Rules 1–4 as a *carry rule*. Moreover, we note that Rules 1 and 2 maintain the sum of digits in the list, while both Rules 3 and 4 reduce the sum of digits in the list by 1, so we refer to Rules 3 and 4 as *drop carries*. In traditional base-b arithmetic, the number of carries when adding the base-b representations of n and m using the standard algorithm is given by $s_b(n)+s_b(m)-s_b(n+m)$, where s_b is the base-b sum-of-digits function. Our definition of drop carries and the Zeckendorf addition algorithm hence provide the following analogous result for the Zeckendorf sum-of-digits function.

Theorem 3.1. *For two natural numbers n and m, the quantity $s_z(n)+s_z(m)-s_z(n+m)$ is the number of drop carries utilized when adding the Zeckendorf representations of n and m.*

We can visualize the two-dimensional sequence of drop carries in a triangular form: entry ℓ in row n of the triangle in Figure 4 shows the number of drop carries when adding the Zeckendorf representations of ℓ and $n-\ell$, which is given by $s(\ell)+s(n-\ell)-s(n)$ according to Theorem 3.1.

With this theorem in place, we turn our attention to defining a new family of "binomial coefficients" that will have the desired property. Given any sequence of

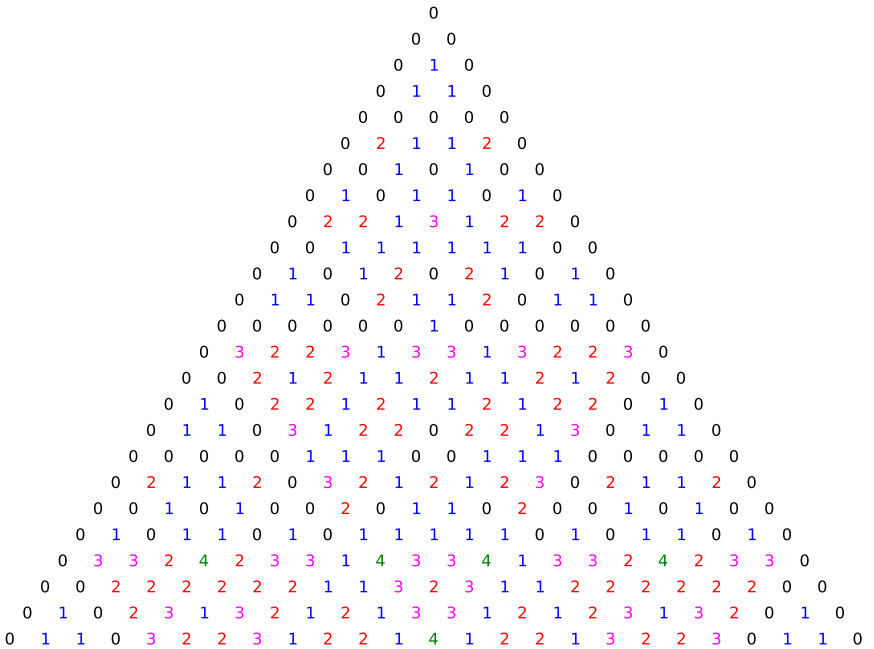

Figure 4. The "drop-carry triangle".

positive integers, $g : \mathbb{N} \to \mathbb{N}_{>0}$, we can define the g-factorial function, $g_!$, to be the sequence of partial products of g:

$$g_!(n) = \prod_{i=1}^{n} g(i).$$

Then, we can use this generalized factorial function to define the generalized binomial coefficients for g, commonly called the g-binomial coefficients:

$$\binom{n}{\ell}_g = \begin{cases} g_!(n)/(g_!(\ell) \cdot g_!(n-\ell)), & \text{if } 0 \le \ell \le n, \\ 0 & \text{otherwise.} \end{cases}$$

Of course, for an arbitrary integer sequence, we cannot expect $\binom{n}{\ell}_g$ to be an integer. We use this construction with the sequence $c : \mathbb{N} \to \mathbb{N}$ defined by

$$c(n) = 2^{\lfloor w(n)/2 \rfloor},$$

whose first few terms are listed in the table below:

n	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
$\lfloor \frac{1}{2} w(n) \rfloor$	0	1	1	0	2	0	1	2	0	1	1	0	3	0	1	1	0	2	0	1
$c(n)$	1	2	2	1	4	1	2	4	1	2	2	1	8	1	2	2	1	4	1	2

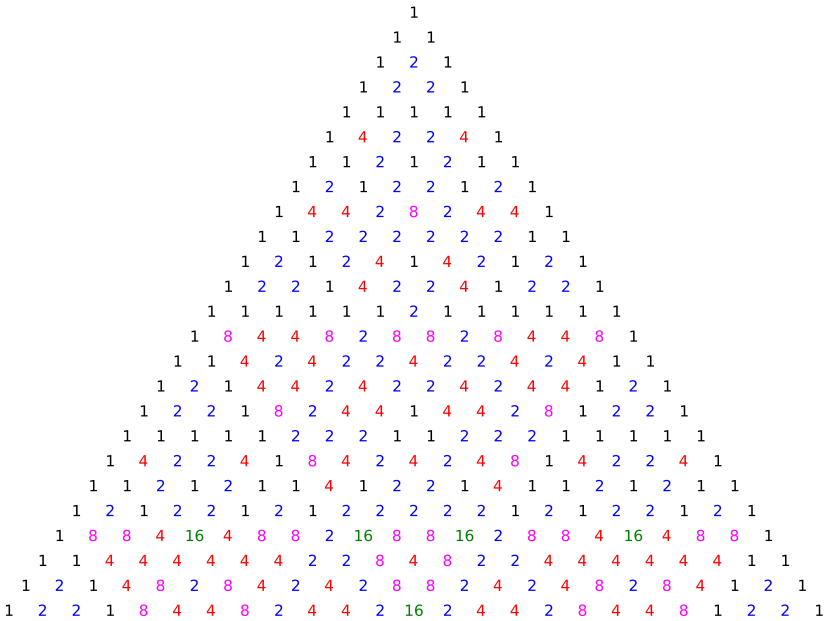

Figure 5. The triangle of c-binomial coefficients.

Knuth and Wilf [1989] showed that if a sequence g is strongly divisible then $\binom{n}{\ell}_g$ will always be an integer, and Edgar and Spivey [2016] showed that if a sequence g is both divisible and multiplicative, then $\binom{n}{\ell}_g$ will always be an integer. Unfortunately, the sequence c is not divisible (note that $c(2) = 2$ and $c(4) = 1$) and thus not strongly divisible. However, it turns out that every c-binomial coefficient $\binom{n}{\ell}_c$ is still an integer.

Theorem 3.2. *Let n and m be natural numbers. Then*

$$\binom{n+m}{n}_c = 2^{s_z(n)+s_z(m)-s_z(n+m)}.$$

Proof. Now, by Theorem 2.7 we know that, for any $j \geq 0$, we have

$$\begin{aligned}
\sum_{i=1}^{j} \left\lfloor \tfrac{1}{2} w(i) \right\rfloor &= \sum_{i=1}^{j} (1 + s_z(i-1) - s_z(i)) \\
&= \sum_{i=1}^{j} 1 + \sum_{i=1}^{j} s_z(i-1) - \sum_{i=1}^{j} s_z(i) \\
&= j - s_z(j) + \sum_{i=1}^{j} s_z(i-1) - \sum_{i=1}^{j-1} s_z(i) \\
&= j - s_z(j) + \sum_{i=2}^{j} s_z(i-1) - \sum_{i=2}^{j} s_z(i-1) = j - s_z(j),
\end{aligned}$$

where the fourth equality follows by reindexing and by noticing that $s_z(0) = 0$. Now, using this fact, we see that

$$\binom{n+m}{n}_c = \frac{2^{\sum_{i=1}^{n+m} \lfloor w(i)/2 \rfloor}}{2^{\sum_{i=1}^{n} \lfloor w(i)/2 \rfloor} \cdot 2^{\sum_{i=1}^{m} \lfloor w(i)/2 \rfloor}} = 2^{\sum_{i=1}^{n+m} \lfloor w(i)/2 \rfloor - \sum_{i=1}^{n} \lfloor w(i)/2 \rfloor - \sum_{i=1}^{m} \lfloor w(i)/2 \rfloor}$$
$$= 2^{(n+m)-s_z(n+m)-(n-s_z(n))-(m-s_z(m))}$$
$$= 2^{s_z(n)+s_z(m)-s_z(n+m)}. \qquad \square$$

Figure 5 shows the triangle of c-binomial coefficients, where the entry ℓ in row n is given by $\binom{n}{\ell}_c$; the previous theorem implies that this triangle contains the same information as the "drop-carry triangle" pictured in Figure 4. Thus, when we put Theorem 3.1 together with Theorem 3.2, we obtain our generalization of Kummer's theorem for Zeckendorf representations.

Corollary 3.3. *Let n and m be natural numbers. Then the exponent of 2 in $\binom{n+m}{n}_c$ is the number of drop carries when adding the Zeckendorf representations of n and m.*

4. Digital dominance, carries and conjectures

Let $n = (n_1, n_2, \ldots, n_k)_z$ and $m = (m_1, m_2, \ldots, m_k)_z$ (where again we append zeroes to each list to ensure they all have the same length.). We say *m Zeckendorf digitally dominates n*, denoted by $n \preceq_z m$, if $n_i \leq m_i$ for all i. This relation $\preceq_z$ is a (lower-finite) partial order on the set of natural numbers (with minimum element 0). Figure 6 provides a visualization of this partial order as a triangular array: entry ℓ

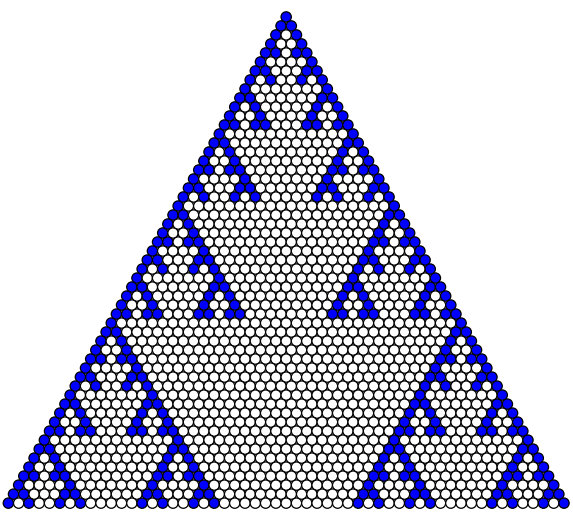

Figure 6. The first 55 rows (starting at 0) of the triangular representation of the Zeckendorf digital dominance order.

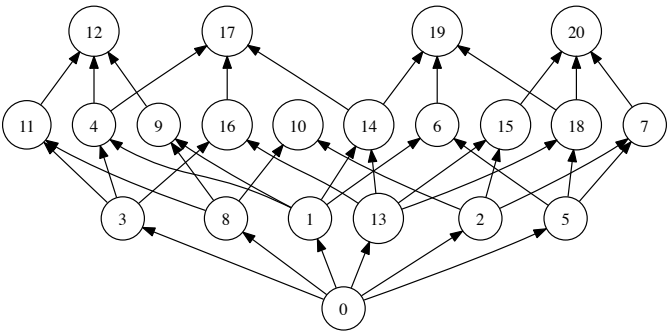

Figure 7. The Hasse diagram of the Zeckendorf digital dominance order up to $n = 20$.

in row n is shaded if and only if $\ell \preceq_z n$. Figure 7 shows the Hasse diagram of the poset $(\mathbb{N}, \preceq_z)$ (up to $n = 20$). This poset is graded with rank function given by the Zeckendorf sum-of-digits.

De Castro et al. [2018] described some connections between the base-b digital dominance order, base-b arithmetic, and binomial coefficients extending some observations by [Ball et al. 2014] related to Fine's theorem [1947] describing how to use Lucas' theorem to count the number of binomial coefficients modulo a prime p. In this section, we introduce some of the same connections and discuss some questions that could be pursued in the future. To begin, we discuss two results about digital dominance.

Proposition 4.1. *Let n and m both be natural numbers with $n = (n_1, n_2, \ldots, n_k)_z$, $m = (m_1, m_2, \ldots, m_k)_z$ and $n + m = ((n+m)_1, (n+m)_2, \ldots, (n+m)_k)_z$, where we append zeroes to each list to ensure they all have the same length*:

(1) *If $n \preceq_z n + m$, then $m \preceq_z n + m$.*

(2) *We have $n \preceq_z n + m$ if and only if $(n+m)_i = n_i + m_i$ for all i.*

Proof. For part (1), let $h_i = (n+m)_i - n_i$. We note that since $n_i \le (n+m)_i \le 1$ for all i, we have $0 \le h_i \le 1$. Then

$$\begin{aligned} h_i \cdot h_{i+1} &= (n+m)_{i+1} \cdot (n+m)_i - (n+m)_{i+1} \cdot n_i - n_{i+1} \cdot (n+m)_i + n_{i+1} \cdot n_i \\ &= -(n+m)_{i+1} \cdot n_i - n_{i+1} \cdot (n+m)_i. \end{aligned}$$

Now, if $(n+m)_i = 0$, then $n_i = 0$ by assumption so that $h_i h_{i+1} = 0$. On the other hand, if $(n+m)_i = 1$, then $(n+m)_{i+1} = 0$ (since we have the Zeckendorf representation), and thus by assumption $n_{i+1} = 0$ since $n_{i+1} \le (n+m)_{i+1}$; hence $h_i \cdot h_{i+1} = 0$. Therefore, $(h_1, h_2, \ldots, h_k)_z$ is a Zeckendorf representation and

moreover

$$\sum_{i=1}^{k} h_i f_i = \sum_{i=1}^{k} ((n+m)_i - n_i) f_i = \sum_{i=1}^{k} (n+m)_i f_i - \sum_{i=1}^{k} n_i f_i = n+m-n = m.$$

Since Zeckendorf representations are unique, we conclude that $h_i = m_i$ for all i. Finally, note that since $n_i \geq 0$ for each i, we have $m_i = h_i = (n+m)_i - n_i \leq (n+m)_i$ for all i, which means that $m \preceq_z n+m$.

For part (2), we first assume that $n \preceq_z n+m$. The result follows by the proof of part (1) since we proved in this situation that $m_i = (n+m)_i - n_i$.

Conversely, if we assume that $(n+m)_i = n_i + m_i$ for all i, then, for each i, we see $n_i \leq n_i + m_i = (n+m)_i$ so that $n \preceq_z n+m$. □

If $n_i + m_i = (n+m)_i$ for all i, then $s_z(n+m) = s_z(n) + s_z(m)$. If this is the case, we say the addition of n and m is *carry-free* since we must only perform the first step of the Zeckendorf addition algorithm to obtain the Zeckendorf representation of $n+m$ from n and m.

Part (1) of Proposition 4.1 explains the symmetry apparent in the digital dominance triangle pictured in Figure 6. Part (2) of the proposition demonstrates a connection between Figure 6 and the drop carry triangle in Figure 4, which is a consequence of the following corollary.

Corollary 4.2. *Let n and m be natural numbers. Then $s_z(n+m) = s_z(n) + s_z(m)$ if and only if $n \preceq_z n+m$.*

In particular, we have that if entry ℓ in row n is shaded in the digital dominance triangle, then entry ℓ in row n of the drop carry triangle is 0 (note that the converse is not true since the digital dominance triangle can detect carries other than drop carries).

The notion of carry-free addition leads us to define a new operation using Zeckendorf representations. For $n = (n_1, n_2, \ldots, n_k)_z$ and $m = (m_1, m_2, \ldots, m_k)_z$ (again with zeroes appended as necessary), we let $n \boxplus m = [n_1 + m_1, n_2 + m_2, \ldots, n_k + m_k]$. Note that $n \boxplus m$ is a list but typically not a Zeckendorf representation since $n \boxplus m$ represents the first list obtained in the Zeckendorf addition algorithm (before utilizing any carry rules). With this notation, we see by Corollary 4.2 that $n \boxplus m$ is the Zeckendorf representation of $n+m$ if and only if $n, m \preceq n+m$. Turning the previous idea around, we fix an integer n and consider the set

$$H_z(n) = \{\ell \boxplus (n-\ell) \mid 0 \leq \ell \leq n\}.$$

We call $H_z(n)$ the set of hyper-Zeckendorf partitions of n. Figure 8 shows the first few values of $h_z(n) := |H_z(n)|$ and then lists them in an irregular table where row y has f_y elements (we start with row 0 and column 1).

The table in Figure 8 has some interesting patterns. In particular, we have the following conjectures.

n	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
$h_z(n)$	1	2	2	2	3	3	3	4	3	5	4	4	5	5	5	6	5	7	6	5

1																				
2																				
2	2																			
3	3	3																		
4	3	5	4	4																
5	5	5	6	5	7	6	5													
7	6	8	7	6	8	8	8	9	7	10	8	7								
9	8	10	10	9	11	10	8	11	9	13	11	10	12	12	11	13	10	14	11	9

Figure 8. The sequence $h_z(n)$ counting the number of hyper-Zeckendorf partitions of n in list form and in irregular table form (where row y has f_y elements).

Conjecture 4.3. *Let $T(n, \ell)$ represent the entry in row n and column ℓ of the irregular table given in Figure 8 (where the first row is 0 and the first column is 1):*

(1) *For all $n \geq 1$, we have $T(n, 1) = T(n, f_n)$.*

(2) *For all n and ℓ with $T(n, \ell)$ defined, we have $T(n, \ell)+T(n+1, \ell) = T(n+3, \ell)$.*

Any formula for $h_z(n)$ would be interesting; part (2) in the conjecture would give a recursion for $h_z(n)$ provided we can find the first three defined values in any column. Part (2) of the conjecture also implies that the column 1 is the Padovan sequence (A000931 in [OEIS]).

Finally, let n be a natural number. For any hyper-Zeckendorf representation $L \in H_z(n)$, we define the set $S(L) \subseteq \{1, 2, 3, \ldots, n\}$ by

$$S(L) = \{\ell \mid \ell \boxplus (n - \ell) = L\}.$$

Now, for any two natural numbers a and b, we let

$$[a, b]_z = \{x \in \mathbb{N} \mid a \preceq_z x \preceq_z b\}$$

and we call $[a, b]_z$ a *dominance interval*. We believe that the set $S(L)$ can be decomposed into dominance intervals.

Conjecture 4.4. *Let n be a natural number and $L \in H_z(n)$. Then there exist integers $a_1, \ldots, a_j$ and $b_1, \ldots, b_j$ such that*

$$S(L) = \bigcup_{i=1}^{j} [a_i, b_i]_z$$

and the union is disjoint.

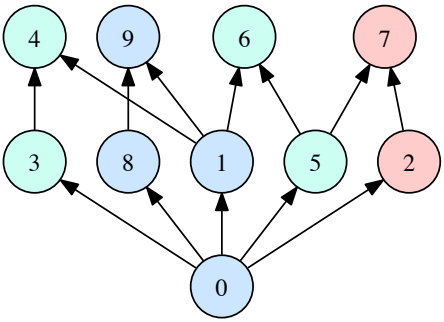

Figure 9. Zeckendorf dominance intervals for $n = 9$.

To demonstrate this conjecture visually, we let $n = 9$ and perform all additions of the form $\ell + (9 - \ell)$:

$$\begin{array}{lll} 0: \quad (0,0,0,0,0)_z & 1: \quad (1,0,0,0,0)_z & 2: \quad (0,1,0,0,0)_z \\ 9: \boxplus (1,0,0,0,1)_z & 8: \boxplus (0,0,0,0,1)_z & 7: \boxplus (0,1,0,1,0)_z \\ \underline{\qquad\qquad} & \underline{\qquad\qquad} & \underline{\qquad\qquad} \\ [1,0,0,0,1] & [1,0,0,0,1] & [0,2,0,1,0] \end{array}$$

$$\begin{array}{ll} 4: \quad (1,0,1,0,0)_z & 3: \quad (0,0,1,0,0)_z \\ 5: \boxplus (0,0,0,1,0)_z & 6: \boxplus (1,0,0,1,0)_z \\ \underline{\qquad\qquad} & \underline{\qquad\qquad} \\ [1,0,1,1,0] & [1,0,1,1,0] \end{array}$$

We see that $h_z(9) = 3$ since $H_z(9) = \{[1,0,0,0,1], [1,0,1,1,0], [0,2,0,1,0]\}$. Furthermore, $S([1,0,0,0,1]) = \{0,1,8,9\}$, $S([1,0,1,1,0]) = \{3,4,5,6\}$ and $S([0,2,0,1,0]) = \{2,7\}$. The Hasse diagram for $\preceq_z$ up to $n = 9$ is pictured in Figure 9.

We see from Figure 9 that $S([1,0,0,0,1]) = [0,9]_z$ (as implied by Corollary 4.2), $S([1,0,1,1,0]) = [3,4]_z \cup [5,6]_z$ and $S([0,2,0,1,0]) = [2,7]_z$. We note that the analogous idea using base-b representations always yields $S(L)$ as a single dominance interval with a constructible minimal element [de Castro et al. 2018]; as such, it would be interesting to know (if Conjecture 4.4 is true) how many intervals are in the union and to find the set of minimal/maximal elements of each dominance interval of which $S(L)$ is the union.

Finally, as we have noted, the algorithm for adding Zeckendorf representations we utilize is not efficient. Additionally, we have forced a particular order in which to perform our rules, but [Fenwick 2003] says that any rule can be used at any time. Theorem 3.1 tells us that we will always have to use the same number of drop carries (regardless of the order we choose to perform rules). Is there some way to determine the minimal number of rules we must use in our Zeckendorf addition? For example,

$$
\begin{aligned}
& & (1,0,1,0,1,0)_z \\
& + & (1,0,1,0,1,0)_z \\
\text{add digits} & \longrightarrow & (2,0,2,0,2,0) \\
\text{Rule 3} & \longrightarrow & (0,1,2,0,2,0) \\
\text{Rule 4} & \longrightarrow & (0,0,1,1,2,0) \\
\text{Rule 4} & \longrightarrow & (0,0,1,0,1,1) \\
\text{Rule 4} & \longrightarrow & (0,0,1,0,0,0,1)_z
\end{aligned}
$$

Figure 10. Addition of Zeckendorf representations: adding $12+12=24$.

Figure 3 required us to use four drop carries and two Rule 1 carries. In Figure 10, we show that if we modify Rule 4 to allow $(\ldots,1,2,0,\ldots)\mapsto(\ldots,0,1,1,\ldots)$, then we can perform the same addition $(12+12=24)$ using only four drop carries and no others.

If $x=n+m$ and $x=(x_1,x_2,\ldots,x_k)_z$, then we can model the carry rules by vectors of length k. The $k-3$ vectors

$$
\begin{aligned}
r_1^1 &:= (1,0,-2,1,\ldots,0),\\
r_1^2 &:= (0,1,0,-2,1,\ldots,0),\\
&\vdots\\
r_1^{k-3} &:= (0,\ldots,0,1,0,-2,1)
\end{aligned}
$$

model Rule 1, the vector $r_2:=(1,-2,1,0,\ldots 0)$ models Rule 2, the vector $r_3:=(-2,1,0,\ldots,0)$ models Rule 3, and the $k-2$ vectors

$$
\begin{aligned}
r_4^1 &:= (1,1,-1,0,\ldots,0),\\
r_4^2 &:= (0,1,1,-1,0,\ldots,0),\\
&\vdots\\
r_4^{k-2} &:= (0,\ldots,0,1,1,-1)
\end{aligned}
$$

model Rule 4. If we let ϵ be the vector with entries given by $\epsilon_i=x_i-n_i-m_i$, then any positive integer solution $(u,y_1,\ldots,y_{k-2},w,z_1,\ldots,z_{k-3})$ to the linear system

$$
ur_3+\sum_{i=1}^{k-2}y_i r_4^i+wr_2+\sum_{i=1}^{k-3}z_i r_1^i=\epsilon
$$

will give us instructions about which rules to apply (though not which order). Therefore, we would be interested in positive integer solutions to the system such that the sum of the entries is minimized. The $k\times(2k-3)$ matrix for this system of

$$M = \begin{bmatrix} -2 & -1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & -1 & -1 & 0 & 0 & 0 & -2 & 0 & 1 & 0 & 0 \\ 0 & 1 & -1 & -1 & 0 & 0 & 1 & -2 & 0 & 1 & 0 \\ 0 & 0 & 1 & -1 & -1 & 0 & 0 & 1 & -2 & 0 & 1 \\ 0 & 0 & 0 & 1 & -1 & -1 & 0 & 0 & 1 & -2 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 1 & -2 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$\operatorname{rref}(M) = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Figure 11. One particular form of the matrix of carry-rule (column) vectors and its reduced row echelon form.

linear equations may thus also be of interest. For instance, Figure 11 shows this matrix when $k = 7$, that is, when the most significant digit of x is x_7. We also include the reduced row echelon form of this matrix and note that if we input the columns systematically, the pattern shown there will continue for all k.

Acknowledgments

This work was performed as part of the 2017 SUMmER REU at Seattle University funded by NSF grant DMS-1460537. We would like to thank Allison Henrich and Steven Klee for their organization of the REU and their guidance during the summer.

References

[Ahlbach et al. 2013] C. Ahlbach, J. Usatine, C. Frougny, and N. Pippenger, "Efficient algorithms for Zeckendorf arithmetic", *Fibonacci Quart.* **51**:3 (2013), 249–255. MR Zbl

[Ball et al. 2014] T. Ball, T. Edgar, and D. Juda, "Dominance orders, generalized binomial coefficients, and Kummer's theorem", *Math. Mag.* **87**:2 (2014), 135–143. MR Zbl

[Benjamin and Quinn 2003] A. T. Benjamin and J. J. Quinn, *Proofs that really count: the art of combinatorial proof*, Dolciani Math. Expositions **27**, Math. Assoc. Amer., Washington, DC, 2003. MR Zbl

[de Castro et al. 2018] P. de Castro, D. Domini, T. Edgar, D. M. Johnson, S. Klee, and R. Sundaresan, "Counting binomial coefficients divisible by a prime power", *Amer. Math. Monthly* **125**:6 (2018), 531–540. MR Zbl

[Edgar and Spivey 2016] T. Edgar and M. Z. Spivey, "Multiplicative functions generalized binomial coefficients, and generalized Catalan numbers", *J. Integer Seq.* **19**:1 (2016), art. id. 16.1.6. MR Zbl

[Edgar et al. 2014] T. Edgar, H. Olafson, and J. Van Alstine, "The combinatorics of rational base representations", preprint, 2014, available at https://community.plu.edu/~edgartj/preprints/basepqarithmetic.pdf.

[Fenwick 2003] P. Fenwick, "Zeckendorf integer arithmetic", *Fibonacci Quart.* **41**:5 (2003), 405–413. MR Zbl

[Fine 1947] N. J. Fine, "Binomial coefficients modulo a prime", *Amer. Math. Monthly* **54** (1947), 589–592. MR Zbl

[Frougny 1991] C. Frougny, "Fibonacci representations and finite automata", *IEEE Trans. Inform. Theory* **37**:2 (1991), 393–399. MR Zbl

[Kimberling 1995] C. Kimberling, "The Zeckendorf array equals the Wythoff array", *Fibonacci Quart.* **33**:1 (1995), 3–8. MR Zbl

[Knuth and Wilf 1989] D. E. Knuth and H. S. Wilf, "The power of a prime that divides a generalized binomial coefficient", *J. Reine Angew. Math.* **396** (1989), 212–219. MR Zbl

[Marsault and Sakarovitch 2014] V. Marsault and J. Sakarovitch, "Breadth-first serialisation of trees and rational languages", pp. 252–259 in *Developments in language theory* (Ekaterinburg, Russia, 2014), edited by A. M. Shur and M. V. Volkov, Lecture Notes in Comput. Sci. **8633**, Springer, 2014. Zbl

[Marsault and Sakarovitch 2017] V. Marsault and J. Sakarovitch, "The signature of rational languages", *Theoret. Comput. Sci.* **658**:A (2017), 216–234. MR Zbl

[OEIS] N. J. A. Sloane et al., "The on-line encyclopedia of integer sequences", available at http://oeis.org.

[Zeckendorf 1972] E. Zeckendorf, "Représentation des nombres naturels par une somme de nombres de Fibonacci ou de nombres de Lucas", *Bull. Soc. Roy. Sci. Liège* **41** (1972), 179–182. MR Zbl

Received: 2019-03-28 Accepted: 2019-06-10

t.ball.6174@gmail.com *Clover Park High School, Lakewood, WA, United States*

rchaiser@gmail.com *University of Colorado Boulder, Boulder, CO, United States*

dtdustin@plymouth.edu *University of Nebraska, Lincoln, NE, United States*

edgartj@plu.edu *Department of Mathematics, Pacific Lutheran University, Tacoma, WA, United States*

pclagarde@mc.edu *South Merrimack Christian Academy, Merrimack, NH, United States*

Guidelines for Authors

Submissions in all mathematical areas are encouraged. All manuscripts accepted for publication in *Involve* are considered publishable in quality journals in their respective fields, and include a minimum of one-third student authorship. Submissions should include substantial faculty input; faculty co-authorship is strongly encouraged. Authors may submit manuscripts in PDF format on-line at the Submission page at the Involve website.

Originality. Submission of a manuscript acknowledges that the manuscript is original and and is not, in whole or in part, published or under consideration for publication elsewhere. It is understood also that the manuscript will not be submitted elsewhere while under consideration for publication in this journal.

Language. Articles in *Involve* are usually in English, but articles written in other languages are welcome.

Required items. A brief abstract of about 150 words or less must be included. It should be self-contained and not make any reference to the bibliography. If the article is not in English, two versions of the abstract must be included, one in the language of the article and one in English. Also required are keywords and subject classifications for the article, and, for each author, postal address, affiliation (if appropriate), and email address.

Format. Authors are encouraged to use LaTeX but submissions in other varieties of TeX, and exceptionally in other formats, are acceptable. Initial uploads should be in PDF format; after the refereeing process we will ask you to submit all source material.

References. Bibliographical references should be complete, including article titles and page ranges. All references in the bibliography should be cited in the text. The use of BibTeX is preferred but not required. Tags will be converted to the house format, however, for submission you may use the format of your choice. Links will be provided to all literature with known web locations and authors are encouraged to provide their own links in addition to those supplied in the editorial process.

Figures. Figures must be of publication quality. After acceptance, you will need to submit the original source files in vector graphics format for all diagrams in your manuscript: vector EPS or vector PDF files are the most useful.

Most drawing and graphing packages (Mathematica, Adobe Illustrator, MATLAB, etc.) allow the user to save files in one of these formats. Make sure that what you are saving is vector graphics and not a bitmap. If you need help, please write to graphics@msp.org with details about how your graphics were generated.

White space. Forced line breaks or page breaks should not be inserted in the document. There is no point in your trying to optimize line and page breaks in the original manuscript. The manuscript will be reformatted to use the journal's preferred fonts and layout.

Proofs. Page proofs will be made available to authors (or to the designated corresponding author) at a Web site in PDF format. Failure to acknowledge the receipt of proofs or to return corrections within the requested deadline may cause publication to be postponed.

involve
2019 vol. 12 no. 7